Volume 5, Issue 4, 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Accuracy Constrained Top down Specialization Approach for Data Anonymization

[1]K. Yasmin Begum, [2]J. Sudha
[1] PG Student, [2] Associate Professor
[1, 2] Department of CSE, A.V.C College of Engineering,
Mayiladuthurai, Tamil Nadu, India

Abstract*— Anonymizing data sets via generality to satisfy certain privacy requirements such as k-anonymity is a widely used category of privacy preserving techniques. Secrecy is one of the most disturbed issues in cloud computing. Personal data like financial operation records and automated health records are extremely sensitive although that can be analyzed and mined by organization. Data privacy issues need to be addressed urgently before data sets are shared on cloud. Data anonymization refers to as hiding complex data for owners of data records. Sharing the private data like economic transaction record in its most specific state poses a threat to individual privacy. Map Reduce algorithm for determining overview and provide protection for sensitive information. Data sets are generalized in a top-down manner until k-anonymity is violated, in order to expose the maximum effectiveness. In this paper, a scalable two-phase top-down specialization (TDS) approach to anonymize extensive data sets using the Map Reduce framework on cloud is to be proposed. In both phases of our approach, we deliberately design a group of innovative Map Reduce jobs to concretely accomplish the concentration computation in a highly scalable way.

***Keywords*— k-anonymity, Data anonymization, Map reduce, Privacy preservation, Sensitive information**

I. INTRODUCTION

Data anonymization is avoiding showing up of sensitive data for holder's data record to mitigate unidentified risk. The privacy of individual can be effectively maintained while some combined information is shared to data user for data analysis and data mining. The proposed method is comprehensive method for data anonymization using Map Reduce on cloud. In First phase, original data set is apportioned into group of smaller dataset and they are anonymized and intermediary result is produced. In second phase, intermediate result first is further anonymized to achieve determined data set. And the data is presented in generalized form using comprehensive approach.

Data anonymization method is used for smacking an identity and of sensitive data for holders of data records. Then, the privacy of individual can be effectively preserved at that time certain accumulation of information is exposed to those data users for diverse analysis and data mining. Data sets scale are important for anonymizing some cloud requests increases very fast in agreement with the cloud computing and Big Data. Data sets have become so large that anonymizing such data sets is becoming a challenge for out-dated anonymization technique. It is important to consent such a system to statement the scalability problem of anonymizing large-scale data set and it is used to give privacy preservation.

The major offerings of this exploration are threefold. First, we creatively apply Map Reduce on cloud to TDS for data anonymization and purposefully design a group of inventive Map Reduce jobs to concretely accomplish the specializations in a highly accessible fashion. Second, a two-phase TDS approach to gain high scalability via allowing specializations to be complemented on multiple data partitions in parallel during the first phase is to be proposed. Third, tentative results show that our approach can considerably improve the scalability and proficiency of TDS for data anonymization over remaining approaches.

II. PROBLEM DEFINITION

An available cloud health service, combinations data from users and shares the data with investigation institutes. Now a days Cloud computing, is become disturbing trend, and it poses a significant impact on current Information Technology industry as well as exploration communities. It provides great scale computation power as well as a loading capacity. A large number of computers together, allowing users to organize applications cost-effectively without heavy infrastructure speculation. Cloud users can decrease huge amount of speculation of IT companies, and deliberate on their own business. The research on cloud privacy and security has come to the picture. Privacy mainly most important issues in cloud computing.

Data privacy can be revealed with less effort by mischievous cloud users or providers because of the failures of some traditional privacy security measures on cloud. Data privacy issues need to be addressed instantly before data sets are investigated or shared on cloud. Data sets have become so large that anonymizing such data sets is becoming a significant challenge for outdated anonymization algorithms. The scholars have begun to investigate the scalability problem of large-scale data anonymization. Addressing the scalability problem of anonymization algorithms via

introducing scalable decision trees and sampling techniques, R-tree index-based approach are used for constructing a spatial index over data sets, achieving high proficiency. However, the above approaches aim at multidimensional generalization only thereby inadequate to work in the TDS approach. As the MapReduce computation prototype is relatively simple, it is still a challenge to design accurate MapReduce jobs for TDS. A data structure Categorization Indexed Partitions (TIPS) is exploited to improve the efficiency of TDS. But the approach is associated, leading to its inadequacy in handling large-scale data sets.

III. PROPOSED WORK

In this proposed work a highly scalable two-phase TDS approach for data anonymization based on MapReduce on cloud is used. The first one, original data sets are partitioned into a collection of smaller data sets, and these data sets are anonymized in parallel, producing transitional results. In the second one, the intermediate results are integrated into one and auxiliary anonymized to achieve consistent k-anonymous data sets. The MapReduce System coordinates the processing by clarifying the scattered servers, executing the various tasks in parallel, protection all communications and data transfers between the various parts of the system, and giving for termination and fault tolerance. A MapReduce program is consists of a Map() procedure that achieves filtering and sorting and a Reduce() procedure that achieves a instantaneous operation .A two-phase TDS approach to gain high scalability via allowing concentrations to be conducted on various data partitions in parallel during the first phase. New approach can expressively progress the scalability and effectiveness of TDS for data anonymization over existing approaches.

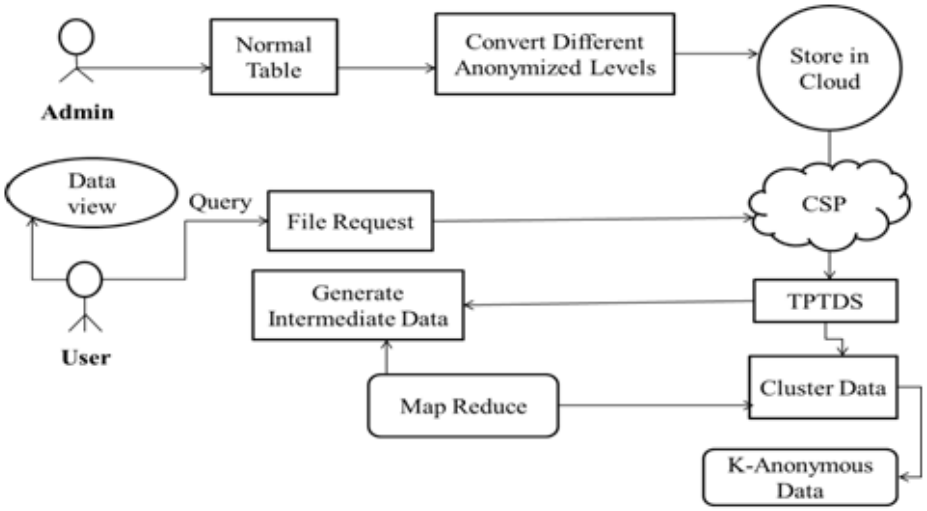

Fig.1 System Architecture

Two-Phase Top-Down Specialization (TPTDS) approach to deportment the computation compulsory in TDS in a highly scalable and proficient fashion. The two parts of the approach are based on the two levels of parallelization provisioned by MapReduce on cloud. Actually, MapReduce on cloud has two levels of parallelization, 1) job level and 2) task level. Job level parallelization means the many MapReduce jobs can be completed instantaneously to make full use of cloud infrastructure resources. Aggregated with cloud, MapReduce becomes more powerful and elastic as cloud can offer substructure resources on request, e.g., Amazon Elastic MapReduce service. Task level parallelization refers to that multiple mapper/reducer tasks in a MapReduce job are executed concurrently over data separations. To accomplish high scalability, parallelizing multiple jobs on data partitions in the first segment, but the ensuing anonymization levels are not related. To obtain definitive consistent unidentified data sets, the second phase is required to incorporate the intermediate outcomes and further anonymize complete data sets.

IV. EVALUATION

A. Micro-Data Creation

Accumulate patient micro data through any hospital or internet and consequence to our application. These data dwelling in Identifiers, Quasi Identifiers, Sensitive Information and this is a normal table ex: Voter Table, Patient Table. The Personal Health Record accumulate different user to our particular web page then that data deposited in public cloud using some anonymization technology.

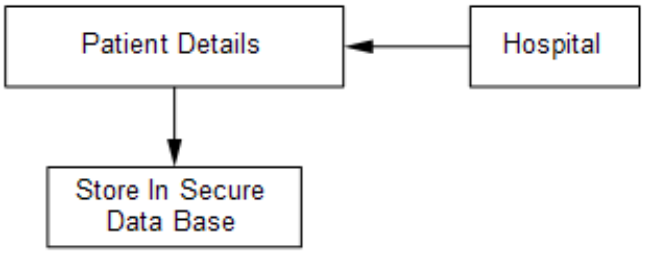

Fig. 2 Micro data creation

B. Anonymization Creation

Health records from a through up hospital located in upstate any country. Note that the table contains no uniquely identifying attributes like name, social security number, etc. Divide the attributes into two groups: the *sensitive* attributes (consisting only of medical condition) and the *non-sensitive* attributes (zip code, age, and nationality). An attribute is marked sensitive if a challenger must not be allowed to determine the value of that attribute for any individual in the dataset.

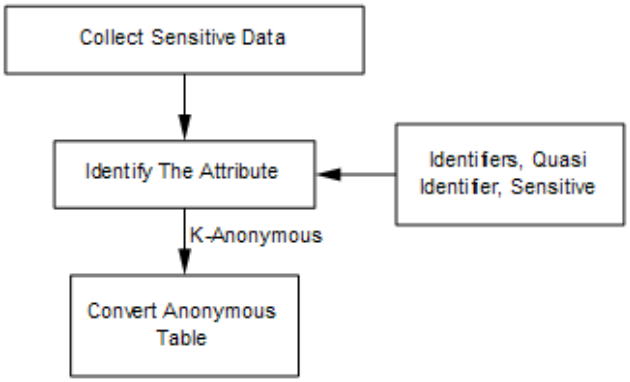

Fig. 3 Anonymization Creation

C. Data Partition

The huge data contains group of private and sensitive data. The each and every sensitive data piercing into small set of data sets. An innovative data set is segregated into group of reduced data sets. The data D is partitioned into $Di,1 \le i \le p,$ it is necessary that the sharing of data records in D_i is similar to D. A data record here can be preserved as a point in an m-dimension space, where m is the number of attributes. Thus, the intermediary Anonymization levels derivative from $Di,1 \le i \le p,$ can be more similar so get a better combined Anonymization level. Arbitrary sampling procedure is adopted to partition D, which can satisfy the above condition. Specifically, an arbitrary number rand $1 \le rand \le p$ is created for each data record. A record is dispersed to the partition D $rand$. Data partition map and reduce algorithm the data partition the number of Reducers should be equivalent to P, so that each Reducer handles one value of rand, exactly producing p resultant files. Each file contains a random sample of D.

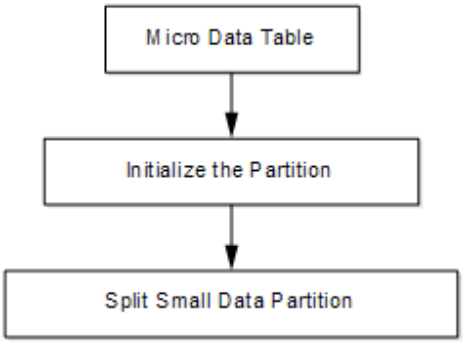

Fig. 4 Data Partition

D. Anonymization level merging

All intermediate Anonymization levels are merged into one in the second phase. The merging of Anonymization levels is completed by merging cuts. Specifically, let Cut_e in AL^l a and Cut_b in AL^l be two cuts of an attribute. There exist domain values $q_a \in Cut_a$ and $q_b \in Cut_b$ that satisfy one of the three conditions: q_a is identical to q_b, q_a is more general than q_b, or q_a is more specific than q_b. To ensure that the merged intermediate Anonymization level AL^l never violates privacy requirements, the more general one is selected as the merged one, for example, q_a will be selected if q_a is more general than or identical to q_b. For the case of multiple Anonymization levels, we can merge them in the same way iteratively. The following lemma ensures that AL^l still complies privacy requirements.

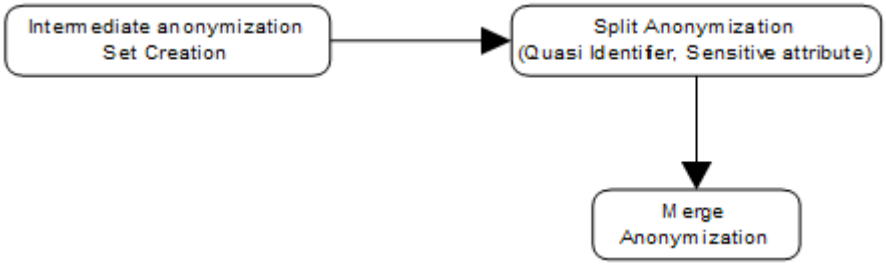

Fig. 5 Anonymization level Merging

E. Data specialization

An original data set D is concretely dedicated for Anonymization in a one-pass Map Reduce job. After finding the combined intermediate Anonymization level $AL^{|}$, run MRTDS (D, k, $AL^{|}$) on the complete data set D, and get the ultimate Anonymization level AL^{*}. Then, the data set D is Anonymize by changing original characteristic values in D with the replying domain values in AL^{*}. Particulars of Map and Reduce functions of the data specialization Map Reduce job are designated in data specialization Map Reduce. The Map function discharges anonymous records and its count. The Reduce function simply combinations these anonymous records and counts their quantity. An anonymous record and its count represent a QI-group. The QI-groups constitute the ending anonymous data sets.

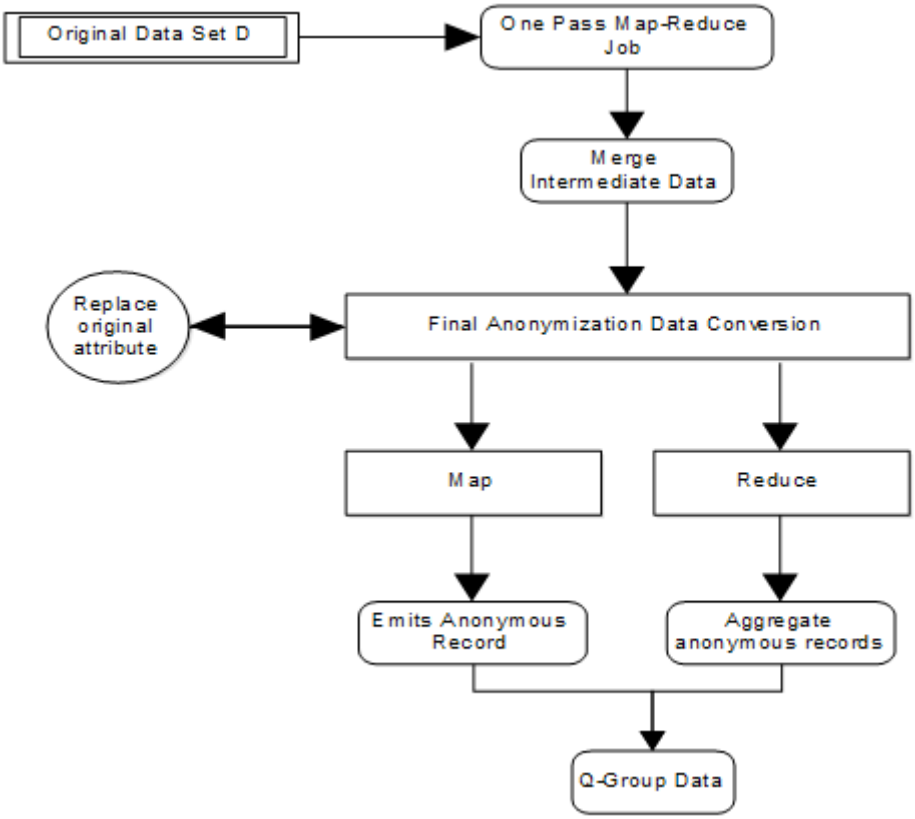

Fig. 6 Data Specialization

F. Map Reduce Two – Phase TDS approach

Map Reduce program consists of Map and Reduce functions, and a Driver that coordinates the command performance of jobs. MRTDS consists of MRTDS driver and two types of jobs. IGPL initialization and IGPL update. The main task of IGPL initialization is to prepare information in the initial Anonymization level AL. The IGPL update job is reasonably similar to IGPl initialization, except that it requires less computation and consumes less network bandwidth.

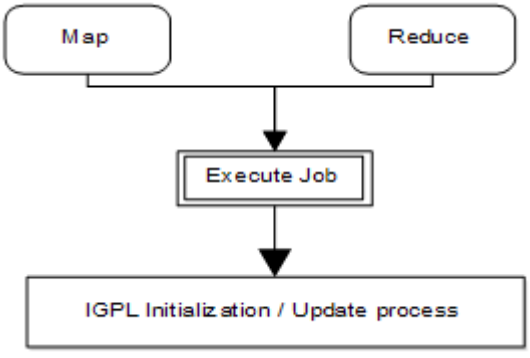

Fig. 7 TDS Approach

V. CONCLUSION

Privacy preserving data investigation and data broadcasting are becoming serious difficulties in today's fragmentary world. In this study we have detected the scalability problematic of large scale data anonymization and found some problems regarding privacy preservation and data gain. To provide these functions a new system is to be proposed a and is similar to large scale data anonymization by TDS approach with privacy preservation and information Gain.The model is stimulated by the map and reducer functions commonly used in programming, even though their

resolution in the MapReduce structure is not the matching as in their original forms. The main influences of the MapReduce framework are not the authentic map and reduce functions, but the extensibility and fault-tolerance extended for a variety of presentations by enhancing the execution engine once.

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] 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.*

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

[5] 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.

[6] 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.

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

