© 2015 IJSRSET | Volume 1 | Issue 1 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

PRESERVING PRIVACY IN CLOUD COMPUTING ENVIRONMENT USING MAP REDUCE TECHNIQUE

Ezhilarasi S[1], Indhumathy R[2], Helen Anitha M[3,] Seetha[4]
Dhanalakshmi College of Engineering, Chennai, Tamil Nadu, India

ABSTRACT

Most of the cloud services require users to share personal data like electronic medical record for data analysis and data mining, bringing privacy concerns. The data sets can be anonymized by using generalization method to attain such privacy requirement. At present the proportion of data in several cloud application growing greatly in congruence with the bid data trend, therefore it is a challenge for currently used software tools to capture , maintain and process such large-scale data within a suffient time. Consequently , it is difficult for achieving privacy due to their inefficiency in handling large scale data sets. In this paper, we propose a scalable two phase top down specialization(TDS) approach to anonymize large-scale data sets using map reduce technique on cloud computing. To achieve the specialization computation in a highly scalable way, we draft a group of creative map reduce jobs in both the phases of our approach. As a result, the experimental evaluation shows that the scalability and efficiency of TDS can be significantly enriched over existing approaches.

Keywords: Data anonymization; top-down specialization; MapReduce; cloud;privacy preservation

I. INTRODUCTION

Cloud computing act as major impact on present IT industry and research communities. Cloud computing provides a great computation power and storage capacity through the use of cast number of computers together, that enable users to deploy their application in a very profitable way. Eventhough, the cloud can reduce the user contribution on IT infrastructure, several customers are still afraid to take use of cloud because of privacy and security issues. Data privacy can be easily thefted by malicious cloud users and providers because of some failures in the commonly used privacy protection techniques on cloud. This can cause substantial economic loss or severe social influence deterioration to data owners, hence data privacy is one of the most concerned issues that need to be addressed imperatively before the data sets are analyzed and shared on cloud.

Data anonymization technique is widely take up for preserving privacy of data that are published and shared among the users on cloud. Data anonymization is the process of hiding identity or more sensitive data that are stored on the cloud. Then the privacy of that data sets can be achieved effectively. For that purpose, a different kind of anonymization algorithms with different anonymization operations have been proposed. Now a days the scale of data sets that are stored and shared between the customers in some cloud application increases rapidly therefore , it is a challenge for existing anonymization algorithms to anonymize such large scale data sets. Map reduce framework have been adopted for handling and processing such large scale-data and that provides greater computation capability for applications. Such frameworks are appropriate to address the scalability problem of anonymizing large scale data sets. Map reduce framework is widely adopted for parallel data processing ,to address the scalability problem of the top-down specialization(TDS) approach for large scale data anonymization.

TDS approach is widely used for data anonymization that provides a good arbitrate between data utility and data consistency. Most of the TDS algorithm are centralized, that are insufficient to handle large-scale data sets. Therefore some distributed algorithms have been proposed. In this paper, we introduce a highly scalable two phase TDS approach for data anonymization by using the map reduce framework on cloud. The anonymization process in the map reduce framework can be divided into two phases. In the first

phase the original data sets splited into smaller group of data sets, and these data sets are anonymized parallely and thus producing a intermediate results. In the second phase the intermediate results are merged together and further anonymize the joined data sets to achieve consistent k-anonymous data sets. A group of map reduce jobs are designed and coordinated to perform specilalization on data sets.

II. METHODS AND MATERIAL

RELATED WORK AND PROBLEM ANALYSIS

2.1 Related Work

The privacy preservation data sets has been broadly studied [1]. Fung et al. [2][3] proposed the TDS approach in that the data sets can be anonymized without the data expedition issue[1].

Mohammed et al. [2] proposed the distributed algorithms to anonymize the large scale data sets that are vertically portioned from various data source without affecting the privacy of data from one consumer to another.

2.2 Problem Analysis

We determine the scalability issue of currently used TDS approaches while managing a vast data on cloud. The centralized TDS approaches in [2][3] uses the TIPS data structure to increase the scalability and efficiency. On the other hand, the quantity of metadata sustained to maintain the statistical information and the linkage information of TIPS data structure is very large compared with the original data sets. Hence , the overhead induced by sustaining the linkage structure and update the statistical data will be large when the data becomes huge.

ELEMENTARY

2.2 Basic Notations

We describe the following underlying representation . Let D indicate a data set comprises of data records. A record r ϵ D has the arrangement r=(v_1, v_2, . . . , v_m , sv), where m is the number of attributes, v_i, $1 \leq i \leq m$, is an attribute value and sv is a sensitive value like interpretation. The set of sensitive values is denoted as SV . An attribute of a record is denoted as Attr, and the taxonomy tree of this attribute is denoted as TT. Let DOM represent the set of all domain values in TT.

2.3 Top-Down Specialization

Typically, TDS is an repetitive approach starting from the primary domain values in the taxonomy trees of attributes. Each round of repetition comprises of three main stages, acquisition of suitable specialization, performing specialization and updating the values of metrics for the next round. The integrity of specialization is deliberated by a search metric. We take over the information gain per privacy loss (IGPL), a compact metric that take up need of both the privacy and the information, as the metric in our TDS approach.

TWO PHASE TOP DOWN SPECIALIZATION (TPTDS)

There are three components in the TPTDS approach; they are data partition, anonymization level merging and data specialization.

2.4 Overview of Two Phase Top-Down Specialization

We suggest a TPTDS approach that are used to perform the computation required in TDS in a highly scalable and efficient way, the two phases of our approach are based on two levels of parallelization produced by map reduce on cloud. There are two levels of parallelization of map reduce on cloud. They are job level and task level. In job level parallelization, the multiple map reduce jobs are executed concurrently. Task level parallelization designate to that different mapper/reducer tasks in a Map Reduce job are executed concurrently over data splits. To acquire high scalability of data sets, we parallelize different jobs on data partitions in the first part , but the outcome of the anonymization levels are not same.

To get the persistent anonymous data sets , it is essential to group the intervening results and anonymize the whole data sets.

ALGORITHM 1 : SKETCH OF TWO- PHASE TDS(TPTDS).

Input: Data set D,anonymity parameters k, k^I and the number partitions p.

Output: Anonymous data set D^*.

1. Partition D into D_i, $1 \le i \le p$.
2. Execute MRTDS (D_i, ^{I}, AL^0) $\rightarrow AL'_i$, $1 \le i \le p$ in parallel as different Map Reduce jobs.
3. Gather all intermediate anonymization level results into one, merge(AL'_1 , AL'_2 ,....., AL'_p)$\rightarrow$ AL^I.
4. Execute MRTDS (D , k, AL^I)AL^* achieve k-anonymity.
5. Specialize D according to AL^*, Output D^*.

2.5 Data Partition

When D is dispersed into D_i, $1 \le i \le p$, it is indispensable that the propagation of data records in D_i is similar to D. A data record here can be appraise as a point in an m-range space, where m is the characteristics. Thus, the moderate anatomy levels imitative from D_i, $1 \le i \le p$, can be more similar so that we can get a good assimilate anatomy level. Random experiment method is embrace to segregation D, which can satisfy the above obligation. Explicitly , a random number rand, $1 \le rand \le p$, is procreated for each data record. A record is empowered to the segregation D_{rand}. Algorithm 2 shows the reproduce program of data partition. Note that the number of Reducers should be equal to p, so that each Reducer handgrip one expense of rand, precisely fertile p backlash files. Each file contains a incidental morsel of D.

ALGORITHM 2: DATA PARTITION MAP& REDUCE

Input : Data record (I Dr, r), r € D, partition parameter p.

Output: D_i, $1 \le i \le p$.

Map: Generate a random number rand, where $1 \le rand \le p$; emit (rand, r).

Reduce: For each rand, emit (null, list(r)).

Once segregated data sets Di, $1 \le i \le p$, are obtained, we run MRTDS (Di, k^I, AL^0) on these data sets in parallel to derive intermediate anonymization levels AL^*_i, $1 \le i \le p$.

2.6 Data Specialization

An genuine data set D is pointedly particular for anatomy in a one-pass MapReduce job. After collecting the assimilate transitional anatomy level AL^I, we run MRTDS (D,k,AL^I) on the integrated data set D, and get the ultimate autonomy level AL^*.

Then, the data set D is anatomized by rehabilitation authenticate peculiarity sphere expense in AL^*. Details of Map and Reduce operations of the data field Map Reduce profession are portray in innovative 3. The Map function emanate anatomy records and its count. The Reduce function quietly accumulates these anatomous documents and poll their number. An anatomous document and its poll denote a QI-group. The QI-groups constitute the ultimate anatomous data sets.

ALGORITHM 3. DATA SPECIALIZATION MAP & REDUCE

Input: Data record (I D_r, r), r € D. ; Anonymization level AL^*.

Output : Anonymous record (r^*, count).

Map: Construct anonymous record $r^* = p_1, (p_2, \ldots, p_m, sv)$, pi, $1 \le i \le m$, 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 $\leftarrow \sum$count; emit (r^*, sum).

MAPREDUCE VERSION OF CENTRALIZED TDS

We complicated the MRTDS in this category. MRTDS performance a nucleus role in two-phase TDS accession, as it is conjured in both aspects to altogether handling calculation. Essentially ,a constructive MapReduce curriculum subsists of map and Reduce objectives, and a motorist that harmonize the large-scale decapitation of assignment.

2.7 MRTDS Driver

Occasionally, an exceptional MapReduce job is depleted to consummate a variegated burden in bounteous germaneness. Thus, an accumulation of MapReduce jobs are integrated in an operator schedule to consummate

such a judicial. MRTDS contains MRTDS Driver and two types of businesses, i.e.., IGPL initialization and IGPL update. The driver arranges the exclusion of jobs.

ALGORITHM 4. MRTDS DRIVER

Input: Data set D, anonymization level AL and k-anonymity parameter k.

Output: Anonymization level AL^I.

1. Initialize the values of search metric IGPL, i.e., for each specialization spec € $U^m_{j=1}$ Cut_j.The IGPL value of spec is computed by job IGPL Initialization.
2. **while** ∃ spec € $U^m_{j=1}Cut_j$ is valid
   a. Find the best specialization from AL_i, specBest.
   b. Update AL_i to AL_{i+1}.
   c. Update information gain of the new specializations in AL_{i+1}, and privacy loss for each specialization via job IGPL Update.
   d. **end while**

   AL'←AL.

III. RESULTS AND DISCUSSION

Implementation and Optimization

To granish how data sets are refined in MRTDS, the execution framework located on standard MapReduce is outlined in Fig. 1. The solid arrow lines shows the data flows in the authorized MapReduce framework. From Fig. 1, we can see that the repetition of MapReduce jobs is controlled by anonymization level AL in Driver. The data flows for handling repetitions are described by dashed arrow lines. AL is accelerated from Driver to all slaves inclusive of Mappers and Reducers via the distributed cache technique. The value of AL is altered in Driver based on the output of the IGPL Initialization or IGPL Update jobs. As the amount of such data is little compared with data sets that will be anonymized, they can be efficiently dispatches between Driver and slaves.

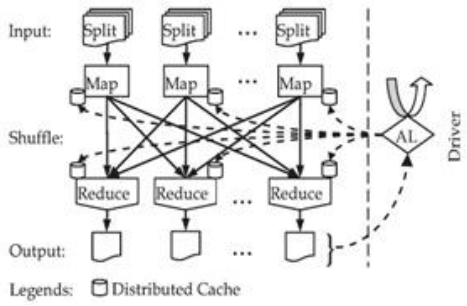

Figure 1 : Execution framework of MRTDS.

IV. CONCLUSION AND FUTURE WORK

In this paper, we have investigated the scalability crisis of large-scale data anonymization by TDS, and anticipated a highly scalable two-phase TDS approach by means of MapReduce on cloud. Data sets are segregated and anonymized simultaneously in the first part, producing intermediate results. Then, the intermediate results are gathered together and further anonymized to fabricate consistent k-anonymous data sets in the second part. We have innovatively applied MapReduce on cloud to data anonymization and intentionally develop a cluster of modern MapReduce jobs to carry out the specialization estimation in a vastly scalable way. Tentative results on real-world data sets have verified that with our approach, the scalability and competency of TDS are enhanced drastically over existing approaches. In cloud , the privacy protection for data analysis, share and mining is a difficult problem because of heavy volume of data sets, thereby requiring rigorous exploration. We will examine the acceptance of our approach to the bottom-up generalization algorithms for data anonymization. Optimized equitable scheduling methods are wonted tto be refined towards complete scalable privacy preservation upon data set scheduling.

V. REFERENCES

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

[2]. 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.

[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]. 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]. 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.

[6]. 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.