International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Data Anonymization Using Map Reduce On Cloud by Using Scalable Two - Phase Top-Down Specialization Approach

Rahul .S Ransing[1], M. S. Patole[2]

[1]Department of Computer Engineering, RMD Sinhgad School of Engineering, Savitribai Phule Pune University, India
[2]Professor, Department of Computer Engineering, RMD Sinhgad School of Engineering, Savitribai Phule Pune University, India

Abstract: ***A cloud services require in big scale, for users to share a private data such as electronic records and health records, transactional data for analysis of data or mining of that data which bringing privacy concerns. We are using k-anonymity concept for the privacy preservation. Recently data in many cloud applications increases in that accordance with the Big Data style, and it make a challenge for commonly used software tools to manage, capture and process on large-scale data within an elapsed time. So, it is a challenge for existing anonymization approaches to achieve privacy preservation on privacy-sensitive large-scale data sets due to their insufficiency of scalability. In this survey paper, we are going to propose and implement a scalable two-phase top-down specialization (TDS) approach to anonymize large-scale data sets using the MapReduce framework on cloud. In both phases of our project, we are going to design a group of inventiveMap Reduce jobs to concretely accomplish the specialization computation in a highly scalable way.***

Keywords: Top Down Specialization, Anonymization of Data, Map Reduce, cloud computing, privacy preservation

1. Introduction

Now a Days Cloud computing, is become disruptive trend, and it poses a significant impact on current Information Technology industry as well as research communities [1] [2]. It provides large scale computation power as well as a storage capacity. A large number of commodity computers together, allowing users to deploy applications cost-effectively without heavy infrastructure investment. Cloud users can reduce huge amount of investment of IT companies, and concentrate on their own business. The research on cloud privacy and security has come to the picture. Privacy mainly most important issues in cloud computing [1].To protect Personal data like electronic health records andfinancial transaction data records are usually deemed extremelysensitive although these data can offer significant humanbenefits if they are analyzed and mined by organizations such as disease research Centre. Data privacy can be unveil with less effort by maliciouscloud users or providers because of the failures of sometraditional privacy protection measures on cloud Data anonymization has been extensively studied andwidely adopted for data privacy preservation in non-interactive process like data publishing and sharing of scenarios[10]. Data anonymization method is used for hiding an identity and of sensitivedata for owners of data records. Then, the privacy of individual can be effectively preserved at that time certain aggregation of information is exposed to those data users for diverseanalysis and data mining.Data sets scale are important for anonymizing some cloud applications increases very fast in accordance with the cloud computing and Big Data [1]. Data sets have become so large thatanonymizing such data sets is becoming a challenge for traditional anonymization technique. It is important to accept such a system to address the scalabilityproblem of anonymizing large-scale data set and it is used to give privacy preservation. In our project we focusMap-Reduce, is widely used for parallel data processing system. Toaddress the scalability problem of the top-down specialization (TDS) approach for large-scale data anonymization[11][12].TDS system, offering a good tradeoff betweendata utility and data consistency, is widely applied for dataanonymization.TDS algorithms are used for centralized, resulting in their lack in handling large-scaledata sets. Although some distributed algorithms have proposed, the main focus on secureanonymization of data sets from multiple parties, ratherthan the scalability aspect. As the Map-Reduce computationprototype is comparatively simple, and still a challenge to designproper Map-Reduce jobs for TDS.In that paper we state a highly scalable two-phaseTDS approach for data anonymization based on Map-Reduce on cloud. To making a full use of the parallel capability ofMap-Reduce on cloud, specializations required. In the process of anonymization it split into the two phases [12]. Firstly original data sets are get partitioned into a group of smaller datasets then these data sets are anonymized in parallel, then it produces intermediate results. Then after, theintermediate results are combined into one, and furtheranonymized to achieve consistent k-anonymous datasets. A group of Map-Reduce jobs are designed and coordinated to perform specializations on data sets collectively. We conclude our methodology by experiments on the real-world data sets. Practically results shows that with our methodology, the scalability and efficiency of TDS can beimproved significantly over existing methods.The major contributions of our research are threefold.First, we apply Map Reduce on cloud to TDS fordata anonymization and deliberately proposed design a group ofinnovative Map Reduce jobs to concretely accomplish thespecializations in a highly scalable manner. Then we state a two-phase TDS approach to get high scalabilityvia allowing specializations to be conducted on multipledata partitions in parallel during the first phase. An experimental result proves that our method can significantly improve the scalability and efficiency of TDS for dataanonymization over existing approaches.

2. Literature Review

Data privacy preservation has been extensively investigated [10]. We briefly review related work below. LeFevre et al

[11].Givesthe scalability problem of anonymization algorithms via introducing scalable decision trees and sampling techniques. Iwuchukwu and Naughton [12] proposed an R-tree index-based approach by building a spatial index on data sets, to achieve high efficiency. Above Explained approaches aim at multidimensional generalization, thereby failing to work in the TDS approach. Fung et al. proposed the TDS approach that produces anonymous data sets without the data exploration problem [10]. A data structure Taxonomy Indexed Partitions (TIPS) is exploited to improve the efficiency of TDS. But the method is centralized, results in lack in handling large-scale data sets.

Several distributed algorithms are proposed to preserve privacy of multiple data sets retained by multiple parties. Jiang and Clifton [13] and Mohammed et al [11]. Proposed distributed algorithms to anonymize vertically partitioned data from different data sources without disclosing privacy information from one party to another. Jurczyk and Xiong [14] and Mohammed et al [11]. Proposed distributed algorithms to anonymize horizontally partitioned data sets maintained by multiple data holders. However, the above Explained distributed methods mainly aim at securely integrating as well as anonymizing multiple data sources. Our work mainly emphasis on the scalability issue of Top downSpecialization anonymization, and is, hence, orthogonal and complementary to them as to Map-Reduce-relevant privacy protection technique, Roy et al [15]. Investigated the data privacy problem caused by Map Reduce and gives a system named Airavat incorporating mandatory access control with differential privacy. Further, Zhang et al [10]. Leveraged Map-Reduce to automatically partition a computing job in terms of data security levelswill help to maintain data privacy in hybrid cloud. Our research exploits Map-Reduce it to anonymize large-scaledata sets before data are further processed by other Map Reduce jobs which are arriving at privacy preservation.

3. Problem Analysis

In this methodwe analyze the scalability problem of existing TDSapproaches when we handling large-scale data sets on cloud platform. In Centralized the TDS approaches [10] [11] it explore the data structure of TIPS to increase the efficiency and scalability by indexing the anonymous data records in data structure

To Increase the specialization process speed because we indexingstructure avoid frequently scanning total data sets andstoring its statistical results. On the other side, the amount of metadata kept as it is to maintain the statistical information and linkage information of the record partitions is relatively large compared withdata sets themselves, thereby consuming assumed memory. Furthermore, the overheads incurred by maintainthe relation of structure and updating the statistic information will be vast when date sets become large. Hence, centralized approaches probably go through from low efficiencyand scalability when handling large-scale data sets.There is a guess that all data processed shouldfit in memory for the centralized approaches [10]. Unfortunately, this guess often fails to hold in mostdata-intensive cloud applications. In the cloudenvironments, computation is provisioned in the form ofvirtual machines. Generally a cloud compute services offer several flavors of VMs. As a result of the centralized approaches are difficult in handling large-scale data setswell on cloud using just one single Virtual Machine even if the VM has the highest computation and storage capacity. A distributed TDS approach[13] is proposed to address the distributed anonymization problem generally concern privacy protection against other parties, ratherthan scalability issue Further, the approach is only to utilizeinformation gain, rather than its combination with privacyloss, as the search metric when we determining the best specializations from this .In a TDS algorithm without considering privacy loss probably chooses a specialization that leads to a sudden violation of anonymity requirements. Hence, the distributed algorithm fails to produce unknown data sets exposing the same data utility as centralized ones this issues like communication protocols and fault tolerance must kept in mind when we designing a distributed technique. As such, it is wrong to control existing distributed technique to compute the scalability problem occurred in the TDS.

Key Terms

3.1 Top-Down Specialization

TDS is repeated process which is starting from the topmost domain values in the arrangement trees of attributes. Each round of iteration consists of 3 main steps. Finding the best specialization, performing specialization and updating values of the search metric for the next round [10]. Such a process of TDS is repeated until k-anonymity isviolated, to description for the maximum data is going to utilize in that. The righteousnessof a specialization is measured by a search metric. In that we acceptthe information gain per privacy loss (IGPL), a tradeoff metric that take in mind both the privacy and information requirements, as the search metric in our approach [11]. Aspecialization with the highest IGPL value is regarded as best one and selected of each round. We describe briefly how we calculate the value of IGPL subsequently to makereaders understand our approach well. Interested readerscan refer to for more details.

3.2 Two-Phase Top-Down Specialization (TPTDS)

There are 3 components present in the TPTDS approach, i.e.
1) Data partition,
2) Anonymization level merging
3) Data specialization

3.2.1 Sketch of Two-Phase Top-Down Specialization

We Gives a TPTDS method to conduct the computation which are required in TDS in a highly scalable and efficient way. The two phases of our method are based on the two levels of parallelization conditioned by Map Reduce on cloud.

Generally Map Reduce on cloud has two levels of parallelization 1] job level and 2] task level. Job level parallelization means that multiple Map-Reduce jobs can be executed concurrently to make a full use of cloud infrastructure resources. Combined with cloud, Map-Reduce become more powerful and stretch as cloud can offer infrastructure resources on require, for example, the Amazon Elastic Map-Reduce service. Task level parallelization is refers to that multiple mapper/reducer tasks in a Map-Reduce job are executed concurrently over data splits. To achieve high scalability, we parallelizing multiple jobs on data partitions in first phase, but the resultant anonymization levels are not same. To obtain finally consistent anonymous data sets, the second phase is important to integrate the intermediate results and further anonymize entire data sets. Details are formulated as follows:

Firstly an original data set D is partitioned into smaller unit .We run a subroutine over each of the partitioned data sets in parallel to make full use of the job level parallelization of MapReduce. The subroutine is a Map Reduce edition of centralized TDS (MRTDS) which concretely conducts the computation is essential in TPTDS. MRTDS anonymizes data partitions is to generate intermediate anonymization levels.

3.2.2 Data Partition

In the Data is partition, Data cut in to number of pieces required that the distribution of data records in Di is similar to D. A data record here can be treating as a point in an m-dimension space,where m is the number of attributes. Random sampling technique is adapted to partition. The number of Reducers should be equal to p, so that each Reducer handles one value of rand, exactly producing p resultant files. Each file contains a randomsample of D.

3.3 Anonymization Level Merging

All middle anonymization levels are merged into one in the second phase. The merging of anonymization levels is completed by merging cuts. For the case ofmultiple anonymization levels, we can merge them in thesame way by iteratively fashion.

3.4 Data Specialization

An original data set D is concretely specialized foranonymization in a one-iteration in Map Reduce job. When we obtainthe merged intermediate anonymization level AL*, we run MRTDS Driver (D, k, AL*) on the entire data set D, and get the final anonymization level AL*.Then Reduce function simply aggregate these anonymous records and counts the number of that particular records. An anonymous record and its count represent a QI-group.

3.5 Map-Reduce Version of Centralized TDS.

In this section we detailed about the MRTDS in thissection. MRTDS Driver plays an important role in the two-phase TDS approach, as it is invoked in these phases to concretely conduct calculation. Basically, practically Map-Reduce program include a Map and Reducefunctions, and a Driver that coordinates the macro execution of jobs came from this stage.

MRTDS Driver Basically, a single Map-Reduce job is insufficient to accomplish a difficult task in many applications. A group ofMap-Reduce jobs are orchestrate in a driver program to achieve such a goal. There are 2 type of jobs in MRTDS Driver i.e., 1] IGPL Initialization and 2] IGPL Update. The MRTDS driver manages an execution process of jobs MRTDS produce the same anonymous data as in the centralized TDS.MTRDS mainly differs from centralized TDS on calculating IGPL values. But, calculating IGPL values dominates the scalability of TDS approaches, as it requires TDS algorithms to count the statistical information of data setsiteratively. MRTDS exploits Map-Reduce on cloud to makethe computation of IGPL parallel and scalable. We presentIGPL Initialization and IGPL Update afterward.

3.5.1IGPL Initialization Job

The important task of IGPL is to initialize the information gain and privacy loss for the all the specializations in the initialanonymization level AL.

In the first, we collect the values for each input key. If a key is forcompute the information gain, then the equivalent statistical information is updated in this Step. Then the reducer just needs to keep statistical information for one specialization ata time, which makes the reduce method which is highly scalable. By this method we initialize our job.

3.5.2 IGPL Job Update

The IGPL Update job dominates the efficiency and the scalability of MRTDS, when it is executed iteratively as given in this method. So far, iterative Map Reduce jobs have not been well supported by the standard Map Reduce framework like Hadoop. Thus, Hadoop variations like Hadoop and Twist have been proposed recently tosupport efficient iterative Map Reduce computation. Our method is based on the standard Map Reduce framework to facilitate the discussion in this.The IGPL Update job is little bit similar to IGPL Initialization, except this it requires less computation and consumes less network bandwidth. Therefore, the former is more efficient than the latter. The Reduce function is the same as the IGPL Initialization, which is already gives in this IGPL Algorithm.

4. Proposed System

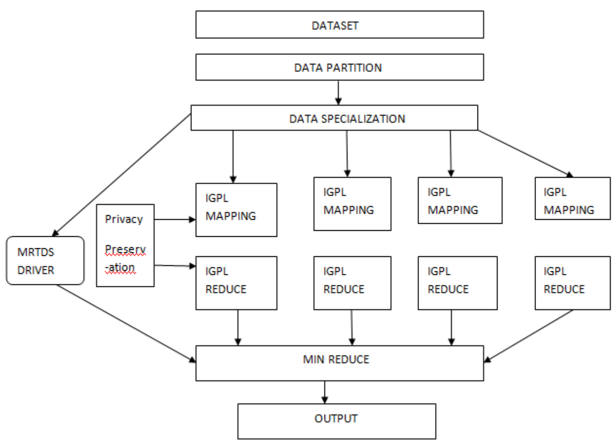

Figure 1: System Architecture

To explain how the data sets are processed in MRTDS, the execution is based on standard Map-Reduce. In that arrow lines represent the dataflows in the Map Reduce framework. From Fig.1, the dataflows for handling iterations are denoted by dashed arrowlines. The value of AL is modified in Driver according to theoutput of the IGPL Initialization as well as IGPL Update jobs. As theamount of such data is small compared with datasets that will be anonymized, and they can be efficiently transmitted between the Driver and workers. We use Hadoop platform, an open-source implementation of Map Reduce, to implement MRTDS Driver. Since most of Map and Reduce functions need to access current anonymization level AL. For this we are going to use the distributed cache mechanism to pass the content of AL to each Mapper or Reducer node as shown in Fig. 1.Hadoop is also provided a mechanism to set simple global variables to Mappers and Reducers. The best specialization result is passed into the Map function of IGPLUpdate job in this way. To minimize the communication traffics, MRTDS exploit combiner mechanism that collect the key-value pairswith the same key into one on the nodes running Map functions. As anonymity computation causes themost traffic as it gives the m key-value pairs for each originalrecord, this can considerably reduce network traffics. Here we are focusing for privacy preservation for this we are also implementing the method for the privacy preservation. By this way we are going to implement our system.

5. Conclusion

In this study we have observed the scalability problem of large scale data anonymization and found some problems regarding privacy preservation and information gain so to provide these functions we have proposed a new system similar to large scale data anonymization by TDS approach with privacy preservation and information Gain.

References

[1] 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 Map Reduce on Cloud", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 25, NO. 2, FEBRUARY 2014.

[2] S.Chaudhuri, "What Next? A Half-Dozen Data ManagementResearch Goals for Big Data and the Cloud," Proc. 31st Symp. Principles of Database Systems (PODS '12), pp. 1-4, 2012.

[3] 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.

[4] L. Wang, J. Zhan, W. Shi, and Y. Liang, "In Cloud, Can ScientificCommunities Benefit from the Economies of Scale?," IEEE Trans.Parallel and Distributed Systems, vol. 23, no. 2, pp.296-303, Feb.2012.

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

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

[7] X. Zhang, C. Liu, S. Nepal, S. Pandey, and J. Chen, "A PrivacyLeakage 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.

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

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

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

[11] 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.

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

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

[14] 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 Security XXIII (DBSec '09), pp. 191-207, 2009.

[15] 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.

Author Profile

Rahul Ransing Research Scholar at RMD Sinhgad School of Engineering, Savitribai Phule Pune University. He has received B.E. in Computer Engineering from University of Pune, Pune. Currently he is pursuing M.E. in Computer Engineering from RMD Sinhgad School of Engineering, Pune, Savitribai Phule Pune University Pune.

Prof. M. S. Patole received the B.E. and M.E. Degrees in Computer Engineering from Savitribai Phule Pune University. She is working as Assistant Professor in Department of Computer Engineering, RMD Sinhgad School of Engineering Pune, India. She is havingmore than six years' experience.