International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 10, October 2015
Available at http://internationaljournalofresearch.org

Two phase top down specialization for high scalability and privacy threads using Map reduce on cloud

Student: S. Maria Jyothi (13H61D0510)
Guide: Mr. Jayendra Kumar (M.Tech) Assistant Professor
(Anurag Group of Institutions)

Abstract—

A large number of cloud services require users to share private data like electronic health records for data analysis or mining, bringing privacy concerns. Anonymizing data sets via generalization to satisfy certain privacy requirements such as kanonymity is a widely used category of privacy preserving techniques. At present, the scale of data in many cloud applications increases tremendously in accordance with the Big Data trend, thereby making it a challenge for commonly used software tools to capture, manage, and process such large-scale data within a tolerable elapsed time. As a result, 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 paper, we propose 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 approach, we deliberately design a group of innovative MapReduce jobs to concretely accomplish the specialization computation in a highly scalable way. Experimental evaluation results demonstrate that with our approach, the scalability and efficiency of TDS can be significantly improved over existing approaches.

Index Terms—Data anonymization; top-down specialization; Map Reduce; cloud; privacy preservation

INTRODUCTION:

CLOUD computing, a disruptive trend at present, poses a significant impact on current IT industry and research communities .Cloud computing provides massive computation power and storage capacity via utilizing a large number of commodity computers together, enabling users to deploy applications cost-effectively without heavy infrastructure investment. Cloud users can reduce huge upfront investment of IT infrastructure, and concentrate on their own core business. However, numerous potential customers are still hesitant to take advantage of cloud due to privacy and security concerns . The research on cloud privacy and security has come to the picture .Privacy is one of the most concerned issues in cloud computing, and the concern aggravates in the context of cloud computing although some privacy issues are not new. Personal data like electronic health records and financial transaction records are usually deemed extremely sensitive although these data can offer significant human benefits if they are analyzed and mined by organizations such as disease research centres. For instance, Microsoft HealthVault ,an online cloud health service, aggregates data from users and shares the data with research institutes. Data privacy can be divulged with less effort by malicious cloud users or providers because of the failures of some traditional privacy protection measures on cloud . This can bring considerable economic loss or severe social reputation impairment to data owners. Hence, data privacy issues need to be addressed urgently before data sets are analyzed or shared on cloud. Data anonymization has been extensively studied and widely adopted for data privacy preservation in noninteractive data publishing and sharing scenarios . Data anonymization refers to hiding identity and/or sensitive data for owners of data records. Then, the privacy of an individual can be effectively preserved while certain aggregate information is exposed to data users for diverse analysis and mining. A variety of anonymization algorithms with different anonymization operations have been proposed. However, the scale of data sets that need anonymizing in some cloud applications increases tremendously in accordance with the cloud computing and Big Data trends [.Data sets have become so large that anonymizing such data sets is becoming a considerable challenge for traditional anonymization algorithms. The researchers have begun to investigate the scalability problem of large-scale data anonymization . Large-scale data processing frameworks like MapReduce have been integrated with cloud to provide powerful computation capability for applications. So, it is promisingto adopt such frameworks to address the scalability problem of anonymizing large-scale data for privacy preservation. In our research, we leverage MapReduce, a widely adopted parallel data processing framework, to address the scalability problem of the top-down specialization (TDS) approach for large-scale data anonymization. The TDS approach, offering a good tradeoff between data utility and data consistency, is widely applied for data anonymization. Most TDS algorithms are centralized, resulting in their inadequacy in handling largescale data sets. Although some distributed algorithms have been proposed . they mainly focus on secure anonymization of data sets from multiple parties, rather than the scalability aspect. As the MapReduce computation paradigm is relatively simple, it is still a challenge to design proper MapReduce jobs for TDS.

In this paper, we propose a highly scalable two-phase TDS approach for data anonymization based on MapReduce on cloud. To make full use of the parallel capability of MapReduce on cloud, specializations required in an anonymization process are split into two phases. In the first one, original data sets are partitioned into a group of smaller data sets, and these data sets are anonymized in parallel, producing intermediate results. In the second one, the intermediate results are integrated into one, and further anonymized to achieve consistent k-anonymous [23] data sets. We leverage MapReduce to accomplish the concrete computation in both phases. A group of MapReduce jobs is deliberately designed and coordinated to perform specializations on data sets collaboratively. We evaluate our approach by conducting experiments on real-world data sets. Experimental results demonstrate that with our approach, the scalability and efficiency of TDS can be improved significantly over existing approaches. The major contributions of our research are threefold. First, we creatively apply MapReduce on cloud to TDS for data anonymization and deliberately design a group of innovative MapReduce jobs to concretely accomplish the specializations in a highly scalable fashion. Second, we propose a two-phase TDS approach to gain high scalability via allowing specializations to be conducted on multiple data partitions in parallel during the first phase. Third, experimental results show that our approach can significantly improve the scalability and efficiency of TDS for data anonymization over existing approaches. The remainder of this paper is organized as follows: The next section reviews related work, and analyzes the scalability problem in existing TDS algorithms. In, we briefly present preliminary for our approach. Section 4 formulates the two-phase TDS approach, and Section 5 elaborates algorithmic details of MapReduce jobs.

We analyze the scalability problem of existing TDS approaches when handling large-scale data sets on cloud. The centralized TDS approaches in and exploits the data structure TIPS to improve the scalability and efficiency by indexing anonymous data records and retaining statistical information in TIPS. The data structure speeds up the specialization process because indexing structure avoids frequently scanning entire data sets and storing statistical results circumvents recomputation overheads. On the other hand, the amount of metadata retained to maintain the statistical information and linkage information of record partitions is relatively large compared with data sets themselves, thereby consuming considerable memory. Moreover, the overheads incurred by maintaining the linkage structure and updating the statistic information will be huge when date sets become large. Hence, centralized approaches probably suffer from low efficiency and scalability when handling large-scale data sets. There is an assumption that all data processed should fit in memory for the centralized approaches .Unfortunately, this assumption often fails to hold in most data-intensive cloud applications nowadays. In cloud environments, computation is provisioned in the form of virtual machines (VMs). Usually, cloud compute services offer several flavors of VMs. As a result, the centralized approaches are difficult in handling large-scale data sets well on cloud using just one single VM even if the VM has the highest computation and storage capability. A distributed TDS approach [20] is proposed to address the distributed anonymization problem which mainly concerns privacy protection against other parties, rather than scalability issues. Further, the approach only employs information gain, rather than its combination with privacy loss, as the search metric when determining the best specializations. As pointed out in [12], a TDS algorithm without considering privacy loss probably chooses a specialization that leads to a quick violation of anonymity requirements. Hence, the distributed algorithm fails to produce anonymous data sets exposing the same data utility as centralized ones. Besides, the issues like communication protocols and fault tolerance must be kept in mind when designing such distributed algorithms. As such, it is inappropriate to leverage existing distributed algorithms to solve the scalability problem of TDS.

Sketch of Two-Phase Top-Down Specialization:

We propose a TPTDS approach to conduct the computation required in TDS in a highly scalable and efficient fashion. The two phases of our approach are based on the two levels of parallelization provisioned by MapReduce on cloud. Basically, MapReduce on cloud has two levels of parallelization, i.e., job level and task level. Job level parallelization means that multiple MapReduce jobs can be executed simultaneously to make full use of cloud infrastructure resources. Combined with cloud, MapReduce becomes more powerful and elastic as cloud can offer infrastructure resources on demand, for example, Amazon Elastic MapReduce service . Task level parallelization refers to that multiple mapper/reducer tasks in a MapReduce job are executed simultaneously over data splits. To achieve high scalability, we parallelizing multiple jobs on data partitions in the first phase, but the resultant anonymization levels are not identical. To obtain finally consistent anonymous data sets, the second phase is necessary to integrate the intermediate results and further anonymize entire data sets. Details are formulated as follows.

IGPL Update Job :

The IGPL Update job dominates the scalability and efficiency of MRTDS, since it is executed iteratively as described in Algorithm 4. So far, iterative MapReduce jobs have not been well supported by standard MapReduce framework like Hadoop [30]. Accordingly, Hadoop variations like Haloop [31] and Twister [32] have been proposed recently to support efficient iterative MapReduce computation. Our approach is based on the standard MapReduce framework to facilitate the discussion herein.

Implementation and Optimization To elaborate how data sets are processed in MRTDS, the execution framework based on standard MapReduce is depicted in Fig. 1. The solid arrow lines represent the data flows in the canonical MapReduce framework. From Fig. 1, we can see that the iteration of MapReduce jobs is controlled by anonymization level AL in Driver. The data flows for handling iterations are denoted by dashed arrow lines. AL is dispatched from Driver to all workers including Mappers and Reducers via the distributed cache mechanism. The value of AL is modified in Driver according to the output of the IGPL Initialization or IGPL Update jobs. As the amount of such data is extremely small compared with data sets that will be anonymized, they can be efficiently transmitted between Driver and workers.

After a specialization spec is selected as the best candidate, it is required to compute the information gain for the new specializations derived from spec. So, Step 1 in Algorithm 7 only emits the key-value pairs for the new specializations, rather than all in Algorithm 5. Note that it is unnecessary to recompute the information gain of other specializations because conducting the selected specialization never affects the information gain of others. Compared with IGPL Initialization, only a part of data is processed and less network bandwidth is consumed. On the contrary, the anonymity values of other specializations will be influenced with high probability because splitting QI-groups according to spec changes the minimality of the smallest QI-group in last round. Therefore, we need to compute AcðspecÞ for all specializations in AL, described in Step 2 and 3 of Algorithm 7. Yet ApðspecÞ can be directly obtained from the statistical information kept by the last best specialization. Note that if the specialization related to pi in Step 3 is not valid, no resultant quasiidentifier will be created.

CONCLUSIONS AND FUTURE WORK:

In this paper, we have investigated the scalability problem of large-scale data anonymization by TDS, and proposed a highly scalable two-phase TDS approach using MapReduce on cloud. Data sets are partitioned and anonymized in parallel in the first phase, producing intermediate results. Then, the intermediate results are merged and further anonymized to produce consistent k-anonymous data sets in the second phase. We have creatively applied MapReduce on cloud to data anonymization and deliberately designed a group of innovative MapReduce jobs to concretely accomplish the specialization computation in a highly scalable way. Experimental results on real-world data sets have demonstrated that with our approach, the scalability and efficiency of TDS are improved significantly over existing approaches. In cloud environment, the privacy preservation for data analysis, share and mining is a challenging research issue due to increasingly larger volumes of data sets, thereby requiring intensive investigation. We will investigate the adoption of our approach to the bottom-up generalization algorithms for data anonymization. Based on the contributions herein, we plan to further explore the next step on scalable privacy preservation aware analysis and scheduling on large-scale data sets. Optimized balanced scheduling strategies are expected to be developed towards overall scalable privacy preservation aware data set scheduling.

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 CostEffective 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] Microsoft HealthVault, http://www.microsoft.com/health/ww/ products/Pages/healthvault.aspx, 2013.