Two-Phase Top-down Specialization for High Scalability and Privacy Concerns

[I]D.S Deva Kiruba Dafi, [II]C.Saravanan, [III]C.Kanimozhi
[I,II,III]VelTech MultiTech DR.Rangarajan & DR.Sakunthala Engineering College, Chennai, India

Abstract

Sharing the private data like financial transaction record in its most specific state poses a threat to individual privacy.MapReduce algorithm for determining generalization 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 utility. This Top-Down Specialization is efficient for high scalability and privacy concerns. High scalable two-phase top-down approach to anonymize large-scale data using mapreduce is proposed.Experimental evaluation result shows that security and privacy preservation of top-down specialization can be significantly improved over existing approach.

Keywords

MapReduce, k-anonymity, Sensitive Information, Data Sets

I. Introduction

Privacy is one of the most concerned issues in cloud computing. Personal data like financial transaction records and electronic 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 sensitive data for owners of data records. Large-scale data sets are generalized using two phase top-down specialization for data anonymization. This process split into two phases. At first mapreduce is applied to the top-down specialization (TDS) and deliberately design a group of innovative mapreduce jobs to accomplish the specialization in high scalable fashion. In second one againthe two-phase top-down specialization is applied to multiple data partition to improve scalability and privacy.

II. Generalization Approach

Data set D contain r number of data records. Each data records have m number of attributes. This attributes are arranged in taxonomy tree structure. The attribute in the taxonomy tree is denoted as TT. Quasi-identifiers QID representing group of anonymous records.

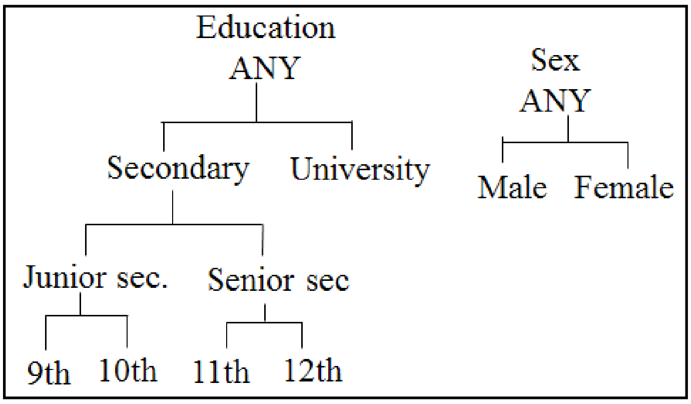


Fig. 1: Taxonomy Tree Structure of Attributes

The top most value of the tree is T. Initially, Cuti contains only the top most value for its attribute. The beneficial specialization in UCuti from the data set to be performed next. At each iteration, find the highest score, denote Best apply to T and update UCuti and update score and validity of the attribute. When Top-Down specialization is applied to the taxonomy tree structure first it Find the best specialization, then perform 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.

III. Large-Scale Data Processing Framework

To address the scalability problem of the Top-Down Specialization (TDS) approach for large scale data set used a widely adopted parallel data processing framework like MapReduce.Mapreduce must have two phases. In first phase, the original datasets are partitioned into group of smaller datasets and these datasets are anonymized in parallel producing intermediate results. In second phase, these intermediate results are integrated into one and further anonymized to achieve consistent k-anonymous dataset.

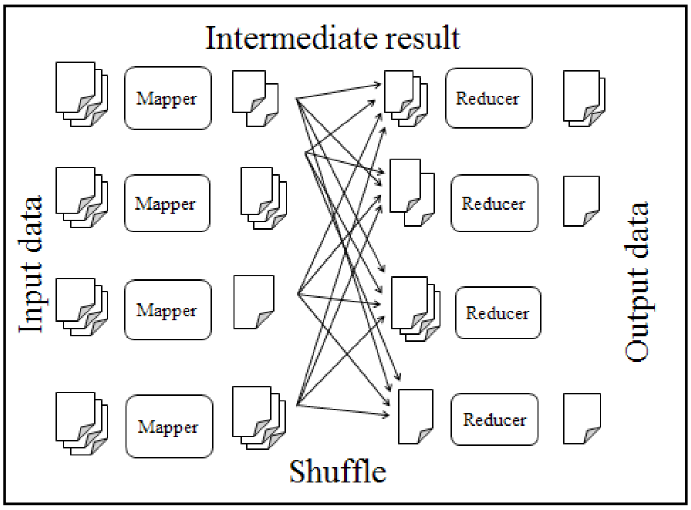
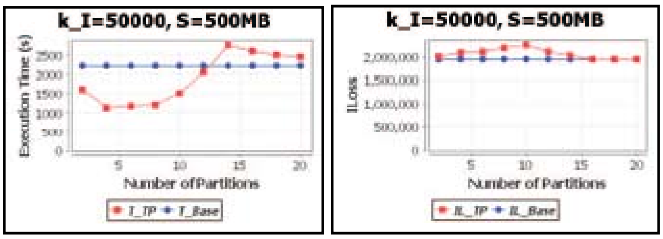


Fig. 2: MapReduce Dataflow

Mapreduce used to splitting up the large input data into chunks of more or equal size, spinning up a number of processing instances for the map phase apportioning data to each of the mappers, tracking the status of each mapper, routing the map results to the reduce phase and finally shutting down the mappers and the reducers when the work has been done. It is easy to scale up MapReduce to handle bigger jobs or to produce results in a shorter time by simply running the job on a larger cluster. When Mapreduce is not used the process fails in distribution system.

IV. Experimental Process

Three groups of experimental in this section to evaluate the effectiveness and efficiency of the approach. In the first one, compared TPTDS with CentTDS from the perspectives of scalability and efficiency. In the other two, investigate on the trade-off between scalability and data utility via adjusting configurations. Generally, the execution time and ILoss are affected by three factors, namely, the size of a data set (S), the number of data partitions (p) and the intermediate anonymity parameter (kI). How the three factors influence the execution time and ILoss of TPTDS is observed. In the first group, measured the change of execution time TCent and TTP with respect to S when p=1. The size S varies from 50 MB to 2.5 GB. In the second group, p is set as 3. The value of p (p>1) is selected randomly and does not affect our analysis as what we want to see is the trend of TTP and ILTP with respect to kI. In the third group, kI is set as 50,000. The value is selected randomly and does not affect our analysis because what we want to see is the trend of TTP and ILTP with respect to the number partitions. The number of partitions varies from 1 to 20.



(a) (b)

Fig. 3: Change in Execution Time and ILoss w.r.t Number of Partitions

V. Conclusion

In this paper, a highly scalable two-phase TDS approach is proposed 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. Experimental results on real-world datasets have demonstrated that with our approach, the scalability and privacy of TDS are improved significantly over existing approach.

References

[1] D. Zissis and D. Lekkas, "Addressing Cloud Computing Security Issues," Fut. Gener. Comput Syst., vol. 28, no. 3, pp 583-592, 2011.

[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," Commun. ACM, vol. 53, no. 4, pp. 50-58, 2010.

[3] X. Zhang, Chang Liu, S. Nepal, S. Pandey and J. Chen, "A Privacy Leakage Upper-Bound Constraint Based Approach for Coat-Effective Privacy Preserving of Intermediate Datasets in Cloud," IEEE Trans. Parallel Distrib, Syst., In Press, 2012.

[4] B. Fung, K. Wang, L. Wang and P.C.K. Hung, "Privacy-Preserving Data Publishing for Cluster Analysis," Data Knowl. Eng., vol. 68, no. 6, pp. 552-575, 2009.

C.BALA SARAVANAN received the M.Tech (IT) from Sathyabama University in 2011. During 2009-2010, I stayed in orbit technologies as software engineer to develop health care automation tool. And I am doing (Ph.D) in VELTECH University and I am now working in VelTech MultiTech Dr.RR & Dr.SR Engineering College as Assistant Professor and IBM TGMC Project coordinator and I published more than 20 journals in varies journal section.

DEVA KIRUBA DAFI D.S pursuing Final Year M.Tech (IT) from VelTech MultiTech Dr.RR & Dr.SR Engineering College.

C.KANIMOZHI pursuing final Year M.Tech (IT) from VelTech MultiTech Dr.RR & Dr.SR Engineering College.