A Scalable Two-Phase Bottom-Up Specialization Prospective for Data Anonymization Using Map Reduce on Cloud

Dilipprasad .E[1], Ajay .R[2], K. Durairaj[3]

Student, Department of Information Technology, Vel Tech University, Chennai, India

Assistant Professor, IT department, Veltech Technical University, Chennai, India

Abstract: ***A scalable two-phase bottom-up specialization prospective to anonymize large-scale data sets using the Map Reduce framework on cloud, this is todays emerging trend and requirement.The people theydon't want to share their sensitive information with unauthorized user, so they want to hide the informationfrom unauthorized person.Emerging techniques such as K-Anonymity, Map Reduceand Data Anonymization are used but there are lot of drawbacksalso occurring in existing method, like privacy, third party access and so on. So in this paperwe are going to see about how bottom up specialization for data anonymization using map reduce on cloud isused for preserving the privacy of user's data's.***

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

1. Introduction

Cloud computing, a disruptive trend at present, poses 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.

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 k-anonymity is a widely used category of privacy preserving techniques. A widely adopted parallel data processing framework, to address the scalability problem of the top down specialization (TDS) prospective for large-scale data anonymization. Most TDS algorithms are centralized, resulting in their inadequacy in handling large-scale data sets. The scalability problem of previous TDS prospective when handling large-scale data sets on cloud. It fails to handles the large amount of the data sets

A scalable two-phase bottom-up specialization prospective to anonymize large-scale data sets using the Map Reduce framework on cloud. Original data sets are partitioned into a group of smaller datasets, and these data sets are anonymized in parallel,producing intermediate results. The Intermediate results are integrated into one, and further anonymized to achieve consistent k-anonymous datasets. A group of Map Reduce jobs is deliberately designed and coordinated to perform specializations on data sets collaboratively.Experimental results demonstrate that with ourprospective, the scalability and efficiency of BUS can beimproved significantly over existing prospective.

The solution, presents a horizontal level of service, available to all implicated Entities, that realizes a security mesh, within which essential trust is maintained. This project recommends introducing a Trusted Third Party, tasked with assuring specific Security characteristics within a cloud environment. The overall performance is good for handling security issues when compare with previous prospective. The quality is best judged with respect to the workload for which the data will ultimately be used.MapReduce allows for distributed processing of the map and reduction operations.

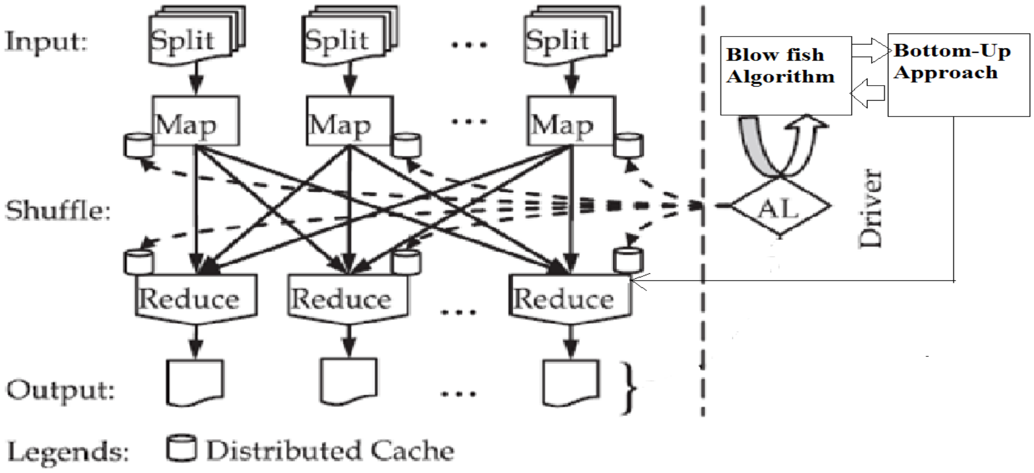
2. Related Work

Recently, data privacy preservation has been extensivelyinvestigated. We briefly review related work below and addressed thescalability problem ofanonymization algorithms via introducing scalable decision trees and sampling techniques. Proposed an R-tree index-based approach bybuilding aspatial index over data sets, achieving high efficiency.However, the above approaches aim at multidimensionalgeneralization thereby failing to work in the TDSapproach. Finally proposed the TDSapproach that produces anonymous data sets without thedata exploration problem. A data structure TaxonomyIndexed Partitions (TIPS) is exploited to improve theefficiency of TDS. But the approach is centralized, leadingto its inadequacy in handling large-scale data sets.Several distributed algorithms are proposed to preserveprivacy of multiple data sets retained by multiple parties. Then proposeddistributed algorithms to anonymize vertically partitioneddata from different data sources without disclosing privacyinformation from one party to another. Proposeddistributedalgorithms to anonymize horizontally partitioned data setsretained by multiple holders. However, the above distributedalgorithms mainly aim at securely integrating andanonymizing multiple data sources. Our research mainlyfocuses on the scalability issue of TDS anonymization, andis, therefore,orthogonal and complementary to them.As to MapReduce-relevant privacy protection, investigated the data privacy problem caused byMapReduce and presented a system named Air vat incorporatingmandatory access control with differentialprivacy. Further, leveraged MapReduceto automatically partition a computing job in terms of datasecurity levels, protecting data privacy in hybrid cloud. Ourresearch exploits MapReduce itself to anonymize large-scaledata sets before data are further processed by otherMapReduce jobs, arriving at privacy preservation.

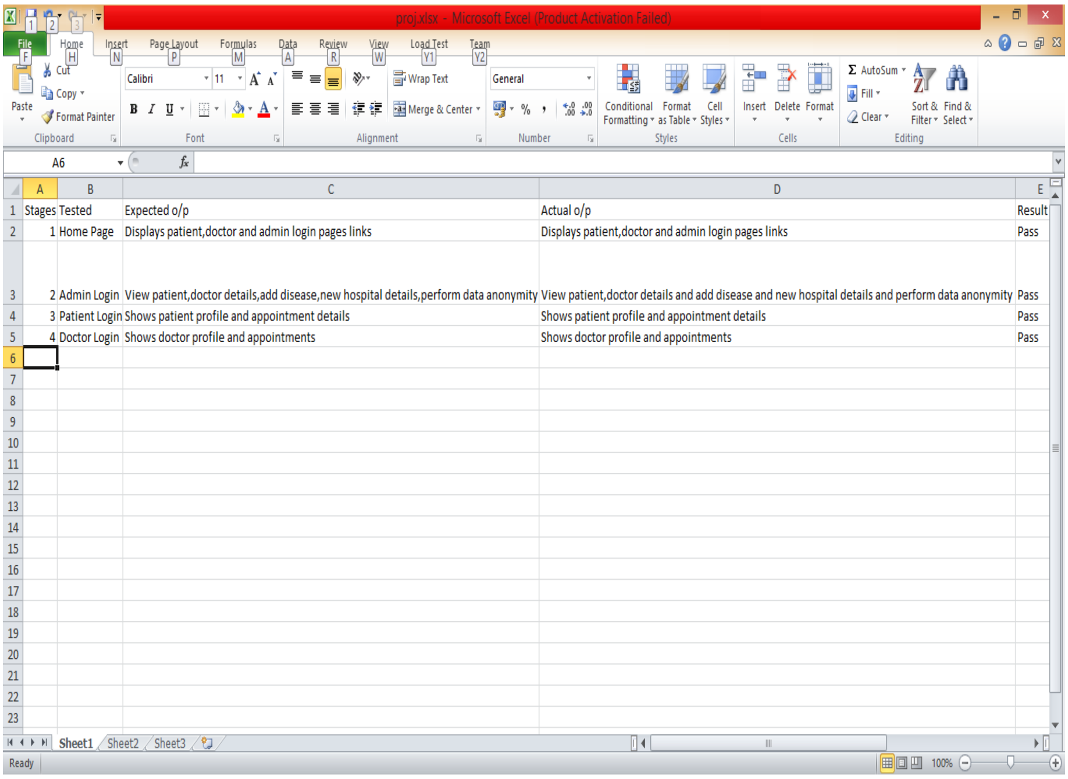
3. Problem Analysis

We analyze the scalability problem of existing TDSapproaches when handling large-scale data sets on cloud.The centralized TDS approaches exploits the data structure TIPS to improve the scalabilityand efficiency by indexing anonymous data records andretaining statistical information in TIPS. The data structurespeeds up the specialization process because indexingstructure avoids frequently scanning entire data sets andstoring statistical results circumvents recompilation overheads.On the other hand, the amount of metadata retainedto maintain the statistical information and linkage informationof record partitions is relatively large compared withdata sets themselves, there by consumingconsiderablememory. Moreover, the overheads incurred by maintainingthe linkage structure and updating the statistic informationwill be huge when date sets become large. Hence,centralized approaches probably suffer from low efficiencyand scalability when handling large-scale data sets.There is an assumption that all data processed shouldfit in memory for the centralizedapproaches assumption often fails to hold in mostdata-intensive cloud applications nowadays. In cloudenvironments, computation is provisioned in the form ofvirtual machines (VMs). Usually, cloud compute servicesoffer several flavours of VMs. As a result, the centralizedapproaches are difficult in handling large-scale data setswell on cloud using just one single VM even if the VM hasthe highest computation and storage capability.A scalable two-phase bottom-up specialization prospective approach is proposed to addressthe distributed anonymization problem which mainlyconcerns privacy protection against other parties, ratherthan scalability issues. Further, the approach only employsinformation gain, rather than its combination with privacyloss, as the search metric when determining the bestspecializations. As pointed out in , a TDS algorithmwithout considering privacy loss probably chooses aspecialization that leads to a quick violation of anonymityrequirements. Hence, the distributed algorithm fails toproduce anonymous data sets exposing the same datautility as centralized ones. Besides, the issues like communicationprotocols and fault tolerance must be kept in mindwhen designing such distributed algorithms. As such, it is inappropriate to leverage existing distributed algorithms tosolve the scalability problem of TDS.

4. System Architecture



5. Performance Analysis



	A	B	C	D	E
1	Stages	Tested	Expected o/p	Actual o/p	Result
2	1	Home Page	Displays patient,doctor and admin login pages links	Displays patient,doctor and admin login pages links	Pass
3	2	Admin Login	View patient,doctor details,add disease,new hospital details,perform data anonymity	View patient,doctor details and add disease and new hospital details and perform data anonymity	Pass
4	3	Patient Login	Shows patient profile and appointment details	Shows patient profile and appointment details	Pass
5	4	Doctor Login	Shows doctor profile and appointments	Shows doctor profile and appointments	Pass

6. Conclusion and Future Enhancement

The investigated the scalability problem of large-scale data anonymization by TDS, and recommend a highly scalable two-phase BUS prospective using Map Reduce 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 Map Reduce on cloud to data anonymization and deliberately designed a group of innovative Map Reduce jobs to concretely accomplish the specialization computation in a highly scalable way. Experimental results on real-world data sets have demonstrated that with our prospective, the scalability and efficiency of BUS are improved significantly over previous prospective.

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. 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 Prospective for Cost-Effective Privacy Preserving of Intermediate Data Sets in Cloud,”IEEE Trans. Parallel and Distributed Systems, to be published, 2012.

[7] 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.