International Journal of Computer Applications Technology and Research
Volume 4– Issue 5, 409 - 413, 2015, ISSN:- 2319–8656

A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using Map Reduce on Cloud

R.Thaayumaanavan
Bharath University
Chennai-600073

J.Balaguru
Bharath University
Chennai-600073

N.Priya
Bharath University
Chennai-600073

Abstract: : More number of users requires cloud services to transfer private data like electronic health records and financial transaction records. A cloud computing services offers several flavors of virtual machines to handle large scale datasets. But centralized approaches are difficult in handling of large datasets. Data anonymization is used for privacy preservation techniques. It is challenged to manage and process such large-scale data within a cloud application. A scalable two-phase top-down specialization (TDS) approach to anonymize large-scale data sets using the Map Reduce framework on cloud. It is used to investigate the scalability problem of large-scale data anonymization techniques. These approaches deliberately design a group of innovative Map Reduce jobs to concretely accomplish the specialization computation in a highly scalable way. The Top-Down Specialization process speeds up the specialization process because indexing structure avoids frequently scanning entire data sets and storing statistical results.

Keywords: map reduce,TDS approach,cloud computing,large scale data set anonymization,privacy preservation,scalable two-phase top-down specialization approach

1. INTRODUCTION

A cloud computing provides efficient computation power and storage capacity via utilizing a large numbers of computers together. On cloud health service, users from various distributed computers can send and share the data in it. Private data like electronic health records or financial transactions are extremely more sensitive if they are used by research centre /accounting entries. They are two conflicting goals that is maximizing data usage and minimizing privacy risk. While determining the best set of transformations has been the focus of extensive work in the database group, the vast majority of this work experienced one or both of the following major problems: scalability and privacy guarantee.

2. EXISTING SYSTEM:

In many cloud applications, data are corrupted in accordance with big data trend while transferring data from one part to another part. At present, we are used software tools like data anonymization via generalization to satisfy certain privacy requirements such as k-anonymity is a widely used category of privacy protecting procedures. Expansive scale datasets have incorporated with cloud applications to provide powerful computation capability. Data anonymization refers to hiding identity and/or sensitive data for owners of data records. Data anonymization approach is used TDS algorithms to handle large scale data sets. It is a challenge to achieve privacy preservation on privacy-sensitive large-scale data sets due to their insufficiency of scalability . Inadequacy inhandling large scale data sets in cloud application.It is failed to achieve high efficiency and File encryption is much difficult

3. PROPOSED SYSTEM:

Data anonymization is difficult in handling of large datasets in cloud applications. It is very challenged to achieve privacy preservation techniques and insufficiency of scalability. To this end, we propose a scalable two-phase top-down specialization (TDS) approach to anonymize large-scale data sets using the Map Reduce framework on cloud. It handles two phase, 1) data partition which describes large datasets are clustered function. 2) Anonymization Level merging which describes clustered tasks are merged into a large-scale dataset. Two- phase TDS combined with Map Reduce framework to reduce unsecured data and maintenance.

Advantages:

It is very easy to access large data set in cloud applications.
The combinations of two-phase TDS, data anonymization and encryption are used in efficient way to handle scalability.

Private data are secured in storage and send transaction forms.

4. SECTIONS
. MODULES:

Login Module:

Administration persons are securely login to our storage management by them via identification and authorization. Only authorized holders enter to view cloud storage information. If the patient wants to view their information or status, they are typing their hospitalization identity number. The users are seen their status when enter the identification number.

Administration Module:

Hospitalization authorities are stored new patient information into the datasets. If the patients are already come into this hospital, update their status via patient treatments. When authorities are want to view some patient information to analyze about health specialization, particular data is required to view full details for preservation of data in cloud applications.

Customer Module:

If other person wants to know particular patient information, patient identification number or hospitalization id must know to view full information. Third parties enter the correct identification number in the customer module. The identification number is validated by login module. If number is correct, patient full details are viewed by the third parties. Customer module is read only module. It does not change or update by patient or any other persons.

Data clustering Module:

Organization persons view all patient information in one selection of administration module. In this module, large datasets are separated by category/department wise. Heart patients are stored separately among all patient information. Likewise, other departments are stored in distinguishable way. The anonymization merge tables are viewed by Data clustering module. The two-phase top down specialization algorithm are applied into the data clustering module to classify each type of category like headache, heart patient, knee department etc. Data are split up into several parts by using first phase of TDS algorithms and data are merged into large datasets by using second type of TDS algorithms. Map is used to data specialization and Reduce framework is used to handle correct dataset organizations.

Privacy Module:

Privacy preservation Techniques are used to data storage applications. Administration authority's data are sometimes easily hacked by malicious users. Overcome of hacker's knowledge, privacy modules use the preservation techniques such as encryption and decryption formats. The special identification of patient is encrypted by authorities and it will store to database in encrypted data. It will decrypt for viewer to read identification is correctly encrypted or not. Two tables are merged to view full details of patient information. In that table, Encrypted identification number is stored replace of original number.

LITERATURE SURVEY:

In the database world, the enterprise data management world, "Big Data" problems arose when enterprises identified a need to create data warehouses to house their historical business data and to run large relational queries over that data for business analysis and reporting purposes. Storage and efficient analysis of such data on "database machines" that could be dedicated to such purposes. Early database machine recommendations involved a mix of novel hardware architectures and designs for prehistoric parallel query processing techniques. Within a few years it became clear that neither brute force scan-based parallelism nor proprietary hardware would become sensible substitutes for good software data structures and algorithms.

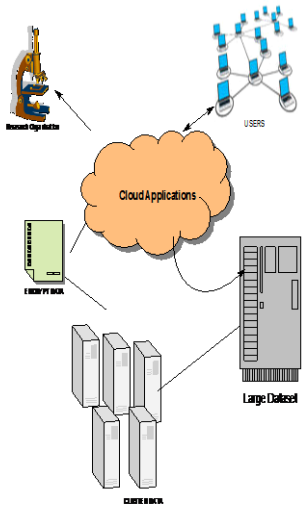
This realization led to the first generation of software-based parallel databases based on the architecture now commonly referred to as "shared-nothing". The architecture of a shared-nothing parallel database system, as the name implies, is based on the use of a networked cluster of Individual machines each with their own private processors, main memories, and disks. All inter-machine coordination and data communication is accomplished via message passing. These

systems exploited the declarative, set-oriented nature of relational query languages and pioneered the use of divide-and-conquer parallelism based on hashing in order to partition data for storage as well as relational operator execution for query processing. A *distributed anonymization protocol* that allows multiple data providers with horizontally partitioned databases to build a virtual anonymize database based on the integration (or union) of the data. As the output of the protocol, each database produces a local anonymize dataset and their union forms a virtual database that is guaranteed to be anonymous based on an anonymization guideline. The convention uses secure multi-party *computation* protocols for sub-operations such that information disclosure between individual databases is minimal during the virtual database construction. L*site-diversity*, to ensure anonymity of data providers in addition to that of data subjects for anonymize data. We present heuristics and adapt existing anonymization algorithms for $l - site - diversity$ so that anonymize data achieve better utility. There are some works focused on data anonymization of distributed databases. presented a two-party framework along with an application that generates *k*-anonymous data from two vertically partitioned sources without disclosing data from one site to the other. Proposed provably private solutions for *k*-anonymization in the distributed scenario by maintaining end-to-end privacy from the original customer data to the final *k*-anonymous results. designing SMC protocols for anonymization that builds virtual anonymize database and query processing that assembles question results. Our disseminated anonymization methodology uses existing secure SMC protocols for subroutines such as computing sum, the *k*-th element and set union. The protocol is carefully designed so that the intermediate information disclosure is minimal. Existing security management and information security life-cycle models significantly change when enterprises adopt distributed computing. Specifically, imparted administration can get to be a significant issue if not legitimately tended to. Regardless of the potential advantages of utilizing clouds, it might mean less coordination among different communities of interest within client organizations. Dependence on external entities can also raise fears about timely responses to security incidents and implementing systematic business continuity and disaster recovery plans. Similarly, risk and cost-benefit issues will need to involve external parties. Customers consequently need to consider newer risks introduced by a perimeter-less environment, such as data leakage within multi-tenant clouds and resiliency issues such as their provider's economic instability and local disasters. Similarly, the possibility of an insider threat is significantly extended when outsourcing data and processes to clouds. Within multi-tenant environments, one tenant could be a highly targeted attack victim, which could significantly affect the other tenants. Existing life-cycle

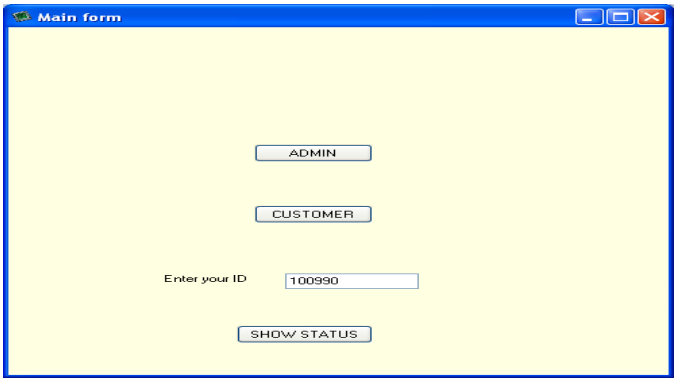
models, risk analysis and management processes, penetration testing, and service attestation must be reevaluated to ensure that clients can enjoy the potential benefits of clouds. The information security area has faced significant problems in establishing appropriate security metrics for consistent and realistic measurements that help risk assessment. We must reevaluate best practices and develop standards to ensure the deployment and adoption of secure clouds. These issues necessitate a well-structured cyber insurance industry, but the global nature of cloud computing makes this prospect extremely complex. Data in the cloud typically resides in a shared environment, but the data owner should have full control over who has the right to use the data and what they are allowed to do with it once they gain access. To provide this data control in the cloud, a standard based heterogeneous data-centric security approach is an essential element that shifts data protection from systems and applications. In this approach, documents must be self-describing and defending regardless of their environments. Cryptographic approaches and usage policy rules must be considered. When someone wants to access data, the system should check its policy rules and reveal it only if the policies are satisfied. Existing cryptographic techniques can be utilized for data security, but privacy protection and outsourced computation need significant attention—both are relatively new research directions.Data provenance issues have just begun to be addressed in the literature. In some cases, information related to a particular hardware component (storage, processing, or communication) must be associated with a piece of data. Although security and privacy services in the cloud can be fine-tuned and managed by experienced groups that can potentially provide efficient security management and threat assessment services, the issues we've discussed here show that existing security and privacy solutions must be critically reevaluated with regard to their appropriateness for clouds. Many enhancements in existing solutions as well as more mature and newer solutions are urgently needed to ensure that

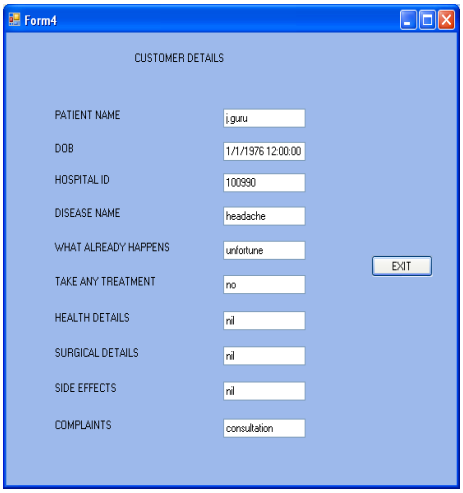
cloud computing benefits are fully realized as its adoption accelerates.

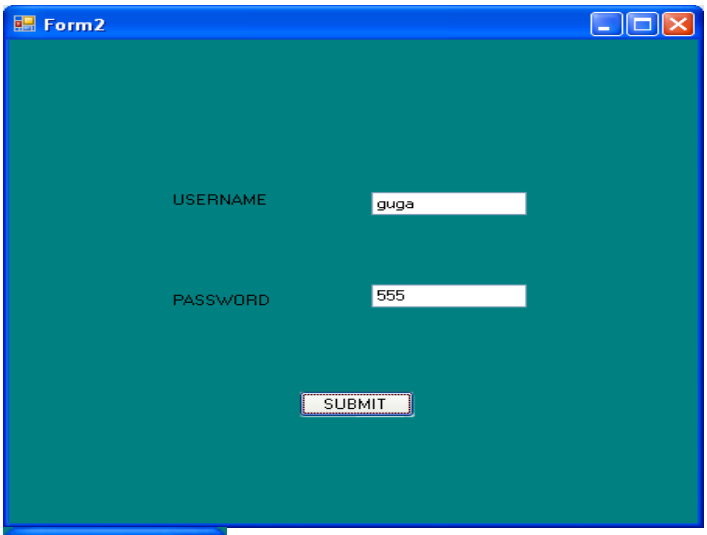
SYSTEM ARCHITECTURE



SCREEN SHOTS:







5. ACKNOWLEDGMENTS

Our thanks to the N.PRIYA(proj guide)&ms.ANURADHA(project coordinator) who have contributed towards development of the template.

6. REFERENCES

[1] S. Chaudhuri, &ldquo,What Next?: A Half-Dozen Data Management Research Goals for Big Data and the Cloud,&rdquo, *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, &ldquo,A View of Cloud Computing,&rdquo, *Comm. ACM,* vol. 53, no. 4, pp. 50-58, 2010.

[3] L. Wang, J. Zhan, W. Shi and Y. Liang, &ldquo,In Cloud, Can Scientific Communities Benefit from the Economies of Scale?,&rdquo, *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, &ldquo,Security and Privacy Challenges in Cloud Computing Environments,&rdquo, *IEEE Security and Privacy,* vol. 8, no. 6, pp. 24-31, Nov. 2010.

.

[5] D. Zissis and D. Lekkas, &ldquo,Addressing Cloud Computing Security Issues,&rdquo, *Future Generation Computer Systems,* vol. 28, no. 3, pp. 583-592, 2011

[6] X. Zhang, C. Liu, S. Nepal, S. Pandey and J. Chen, &ldquo,A Privacy Leakage Upper-Bound Constraint Based Approach for Cost-Effective Privacy Preserving of Intermediate Data Sets in Cloud,&rdquo, *IEEE Trans. Parallel and Distributed Systems,* to be published, 2012.

[7] L. Hsiao-Ying and W.G. Tzeng, &ldquo,A Secure Erasure Code-Based Cloud Storage System with Secure Data Forwarding,&rdquo, *IEEE Trans. Parallel and Distributed Systems,* vol. 23, no. 6, pp. 995-1003, 2012.

.