European Journal of Applied Sciences 8 (3): 186-192, 2016
ISSN 2079-2077

DOI: 10.5829/idosi.ejas.2016.8.3.23005

Data Anonymization of Vertically Partitioned Data Using Mapreduce on Cloud

M. Rajesh kumar and J. Jesila

Master of computer Application, Sathyabama University, India

Abstract: In the world of computers, cloud services, on large scale, are being offered by service providers. User wishes to share some private information that has been stored on the cloud server due to various reasons such as data analysis, data mining and so on. These things bring up a concern about privacy. Privacy preservation may be attained by Anonymization data sets via normalization for satisfying privacy needs by making use of k-anonymity method that happens to be one widely employed kind among the privacy preserving methods. During the current period, data on cloud applications have greatly been found to increase their scale progressively in connection with the trend of Big Data. Therefore it becomes really difficult to manage, accept, process and maintain such huge volume of data within the stipulated time stamps. Hence, it is a very tough task preserving privacy of sensitive and huge sized data. Privacy preservation using existing anonymization methods may not prove efficient because they are not able to handle the scaled datasets. The approach handles anonymization issue on very huge scale cloud datasets by making use of two phase top down specialization method and MapReduce framework. Novel MapReduce tasks are cautiously devised in both the stages of this method for achieving specialization calculation on datasets that are scalable. Efficiency and scalability of the Top down Specialization (TDS) is found to increase significantly over the presently existing method.

Key words: Top Down Specialization · MapReduce · Data Anonymization · Cloud Computing · Privacy Preservation

INTRODUCTION

Cloud computing happens to be one rapidly growing and disruptive tendency during the present times. It constitutes a vital impact on current IT industry and research communities [1]. Cloud computing offers huge power of computation and huge storage capability through the utilization of an increased number of different computer systems working together and thus empowers people to set out many applications in a cost-effective way. Users are not required to invest heavily on infrastructure toward simple computations. In the early period, people had to invest on a large manner for massive computation that is needed occasionally and this does not prove cost-effective. Whereas the cloud service suppliers maintain large infrastructure which the different users may use when and as they need (pay when you grow) without having to invest effort and money on setting up the infrastructure. Nonetheless, there are several people who do not wish to take advantage of the cloud due to security and privacy grounds [2-8]. The research work that has been conducted on cloud security and privacy is coming to picture now [2]. Privacy is considered as one among the most crucial factors in cloud computing. The problem concerned aggravates among cloud environment where privacy-related concern [6] is not new. Financial and health transaction records related to people are considered very sensitive. It is possible for humans to benefit from sensitive information when such data are examined and get mined properly through certain organizations such as disease analysis labs. For instance, consider MicroSoft Health Vault. Hey may access the data, aggregate the same and finally share it with research centers. It is possible to disturb data privacy with least amount of efforts by the cloud users during the time of old privacy preservation systems. This may tend to spoil

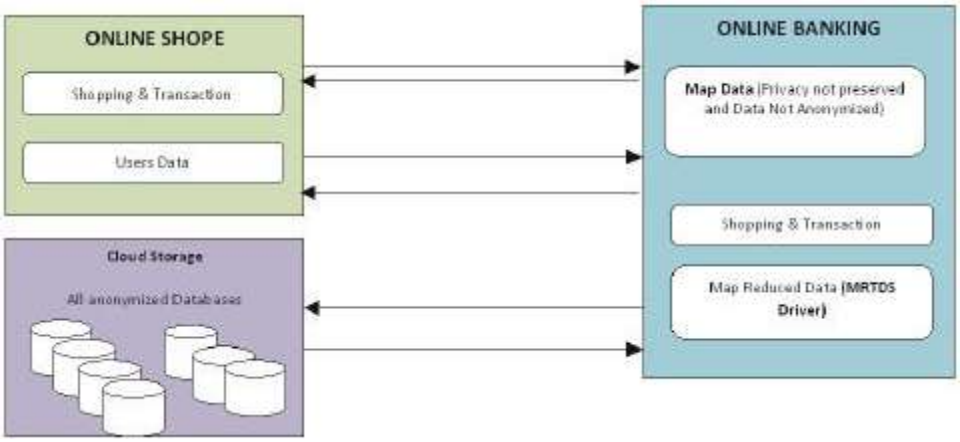
Corresponding Author: M. Rajesh kumar, Master of computer Application, Sathyabama University, India.

the social and financial values of the public. Therefore, it becomes crucial to resolve the privacy problem of information on an urgent basis in cloud environment before they subsequently are being shared among users on the cloud. Anonym zing information is one method that has widely been examined and employed due to the information privacy preservation among the non-interactive data sharing and publishing environments [9]. Data anonymization happens to be one method wherein sensitive information and identity about the data owner's records are hidden. User will be able to utilize the information regarding several reasons such as mining and diverse analysis, thus necessitating the requirement of privacy preservation of datasets on cloud. There are several anonymization algorithms (methods) making use of various anonymization processes that have been suggested [10-13]. Along with time, datasets scales keep on rapidly increasing over the concern about Big Data and also cloud computing which requires anonym zing in certain application of the cloud. Conventional algorithms regarding anonymization happen to be difficult tasks with respect to anonymization of cloud datasets.

Related Work: An information Taxonomy Indexed Segregation structure (TIPS) has been exploited for improving the effectiveness of TDS. Yet, this method is normally centralized, rendering it inadequate for managing huge-scale datasets. Many distributed algorithms have been suggested for privacy preservation with regard to multiple datasets that are retained by several clients. A distributed algorithm was then suggested for anonym zing vertically divided information from various information sources without having to disclose privacy data from one client to the other. It was then suggested to employ distributed algorithms for anonym zing horizontally divided datasets that are retained by several holders. Nevertheless, the primary aim of the above mentioned distributed algorithms was securely anonym zing and integrating multiple information sources. Our study primarily focuses on scalability problem regarding TDS anonymization and hence, it is complementary and orthogonal to them. Regarding Map Reduce-related privacy preservation, we have investigated the information privacy issue that is caused due to Map Reduce and furnished one system called Air vat integrating conditional access control along with varied privacy. Moreover, we have leveraged Map Reduce for automatically partitioning a given computing task in connection with information security levels, preserving data privacy among the hybrid cloud. The suggested analysis tends to exploit Map Reduce itself for anonymizing huge-scale datasets prior to data being processed further by some other Map Reduce tasks that arrive at privacy preservation. Information privacy preservation is extensively seen to be examined [4]. We are reviewing the related work briefly hereunder. LeFevre *et al*. [11] analyzes the scalability issue regarding anonymization algorithms through the introduction of certain sampling methods and climbable decision trees. Naughton and Iwuchukwu suggested an R-tree index-oriented strategy by constructing one spatial index toward the datasets for achieving great efficiency. The aim of the approaches explained above is multi-dimensional normalization, thereby getting defeated in working in TDS method. Fung *et al*. put forward the TDS method which creates anonymous datasets that does not involve information exploration issue [4]. An information structure Taxonomy Indexed Partitions (TIPS) is being exploited for enhancing the effectiveness of TDS. Yet, this strategy happens to be centralized, resulting in lack of efficiency to handle huge-scale datasets. Many distributed algorithms have been suggested for preserving privacy of multitudinous datasets that have been retained by several parties. Mohammed *et al* [11] and [14] Jiang and Clifton introduced distributed algorithms for anonym zing vertically divided information from various information sources without having to disclose privacy data from one client to the other.

Overall Architecture: We inventively apply Map Reduce in the cloud on to TDS pertaining to data anonymization and purposely devise a set of creative Map reduce tasks for accomplishing concretely the specializations with greatly scalable fashion. Next, secondly, we introduce a two-phase TDS strategy for gaining great scalability by permitting specializations that need to be carried on over multiple information segregations in parallel during the initial stage. Thirdly, results from experiments prove that our proposed method will be capable of significantly improving efficiency and scalability of TDS toward data anonymization above the existing methods [13-15].

TWO-PHASE TOP-DOWN SPECIALIZATION (TPTDS)



In the process of Banking, any user has got to register toward enabling online banking facility. After registration, user will be able for logging in through using his or her credentials. When successful logging in is completed, users will be able to create transaction password. Then users can transfer their funds to any other account. In this process, users are capable of changing their login and profile password at any point of time. We have applied an information partition pertaining to the user information during the process of their registration. In the process of online shopping, user will have to first register online. It is possible for the user to login after successfully registering himself or herself. Users get to select their desired products and also add these products on to the shopping cart. Now users will have to validate their request for payment and after conformation of payment, the given page gets re-directed to login page of their Bank. Here, the users have to enter their login details of the bank and validate their deal by keying in their transaction logon credentials. The history of the transaction is being recorded in the banks website. After the deal has been successfully completed, the page gets redirected again to the website of online shopping. The Map Reduce in cloud to TDS toward data anonymization and purposely devise a set of creative Map reduce tasks for accomplishing concretely the specializations with great scalable fashion. For achieving high scalability, multiple tasks have been parallelized on data divisions in first stage, but resultant levels of anonymization will not be identical. For obtaining the final harmonious anonymous datasets, it is essential for the second stage to fuse the transitional results and also further anonymize the entire datasets. Normalization of the information is then applied here. Administrator present in online purchase site will attempt retrieving the information by making use of web service. Information are then displayed in generalized format.

Two Phase Top down Specialization: Two Phase Top Down Specialization (TPTDS) method for conducting computation is needed in TDS on an efficient and greatly scalable fashion. The methods two phases are founded on two ranks of parallelization provided by MapReduce in the cloud. Fundamentally, MapReduce on the cloud will have two ranks of parallelization, namely, task level and job level. Job rank parallelization process refers to multiple Map Reduce-related jobs may be performed at the same time for utilizing the resources of cloud framework to the fullest extent. In combination with the cloud, MapReduce gets more elastic and robust since cloud offers resources of infrastructure on request (for example, Amazon Flexible MapReduce amenity). Task-rank parallelization states that several reducer/mapper tasks present in one MapReduce job can be performed at the same time over information splits. For achieving great scalability, parallelization of several jobs on information divisions during the first stage, but the output anonymization ranks may not be identical. For obtaining final constant anonymous datasets, second stage is required to combine the transitional outputs and also further anonymize the complete datasets. Details can be formulated as under. Every transitional anonymization ranks get merged as one during the second stage. Merging the anonymization ranks will be completed via merging cuts. Particularly, let in between and in two cuts belonging to an attribute. Domain values which will satisfy in general than more

particularly will exist. For ensuring that merged transitional anonymization rank does not violate privacy needs ever, the more common of them is chosen to be the merged one. For instance, it will be chosen when it is identical or more general to the other. Regarding multiple anonymization ranks, it will merge them iteratively in very same manner. Following lemma will assure that data will still comply with privacy needs.

Mapreduce: Map Reduce is one programming paradigm regarding processing of huge datasets using a distributed, parallel algorithm on any cluster. And a MapReduce scheme is constructed of Map () process which carries out sorting and filtering (like sorting employees by last name in queues, with one queue in connection with every name) and one Reduce () process that carries out a cumulative function (like counting number of employees in every queue, giving out frequencies of names). MAP Reduce gets orchestrated through carrying out different tasks parellely, marshalling of distributed servers, handling all data transfers and communications between the different parts of given scheme, taking care of redundancy and error tolerance and general handling of the entire procedure. Inspired by map, this model reduces the functions that are normally being used in the functional programming, albeit that fact that their main purpose in MapReduce framework does not happen to be the same like their authentic forms.

Algorithm: Mrtds

Input: DS- Data set AP- anonymized point and K-parameter of anonymity

Output: AP-Anonymization

Step 1: Initiate value of IGPL search metrics, i.e., for every specialty spec $\in_{q_j^m=1}$ cut_j

IGPL initialization job calculated with the help of IGPL value

Step2 : $\in$ spec $\in_{q_j^m=1}$ cut_j is valid

Step2.1: From AP_i spec find specialty

Step2.2: update AP_{i+1} and AL_i

Step2.3: update Al_{i+1} gained information also update IGPL privacy

Step3: end while AP'← AP

RESULT AND DISCUSSION

The Figure2 shows banking registration. The users have registers their personal details, contact details and login details to make online transaction.

The Figure3 shows transaction credentials creation. In order to make transaction the users have to create user id and password for transaction.

The Figure3 shows transaction. To transfer the amount from one account to another account the users have to give account holder name, account number and amount to transfer.

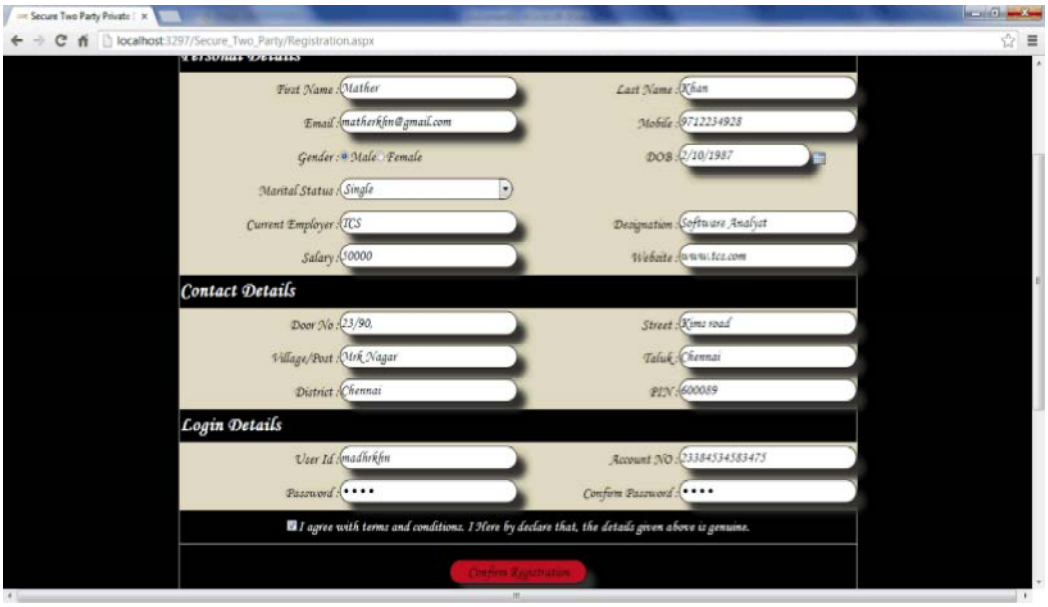


Fig. 2: Banking Registration

Fig. 3: Transaction Credentials Creation

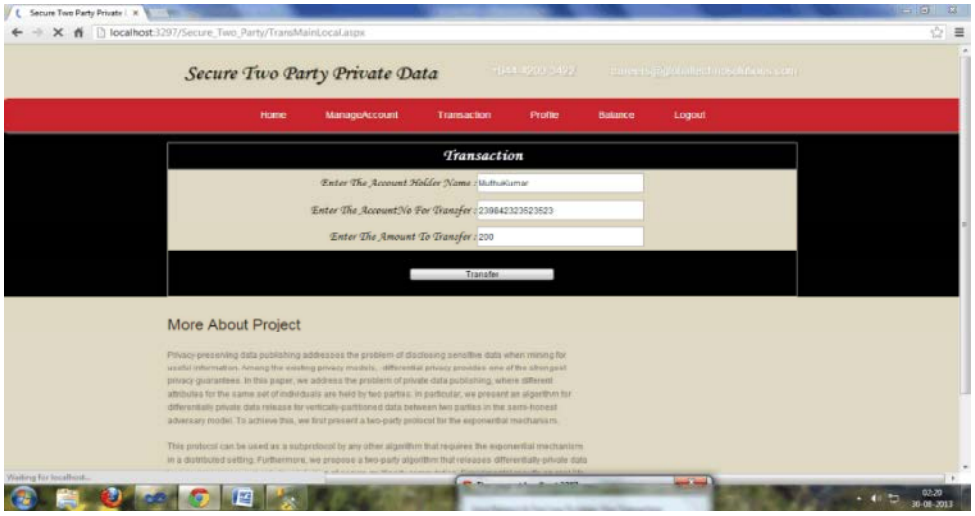


Fig. 4: Transaction

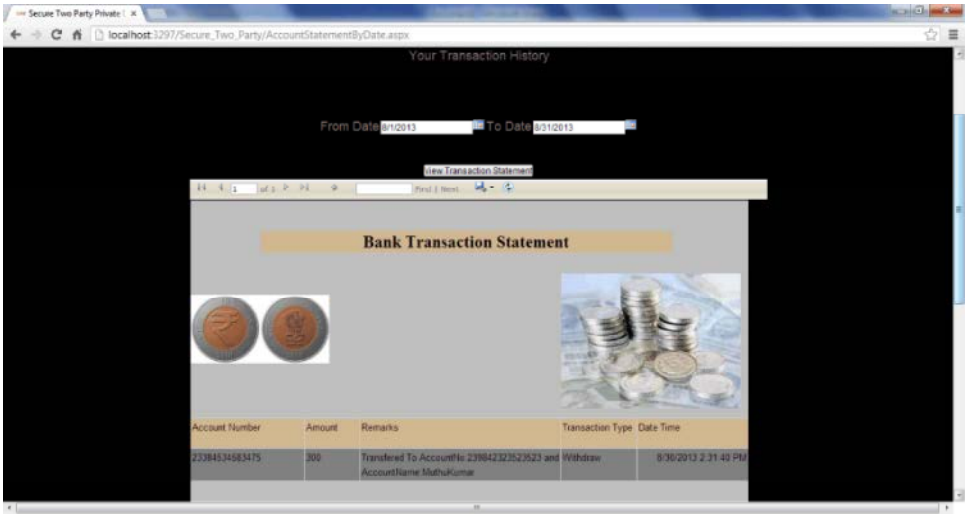


Fig. 5: Transaction Through the Date

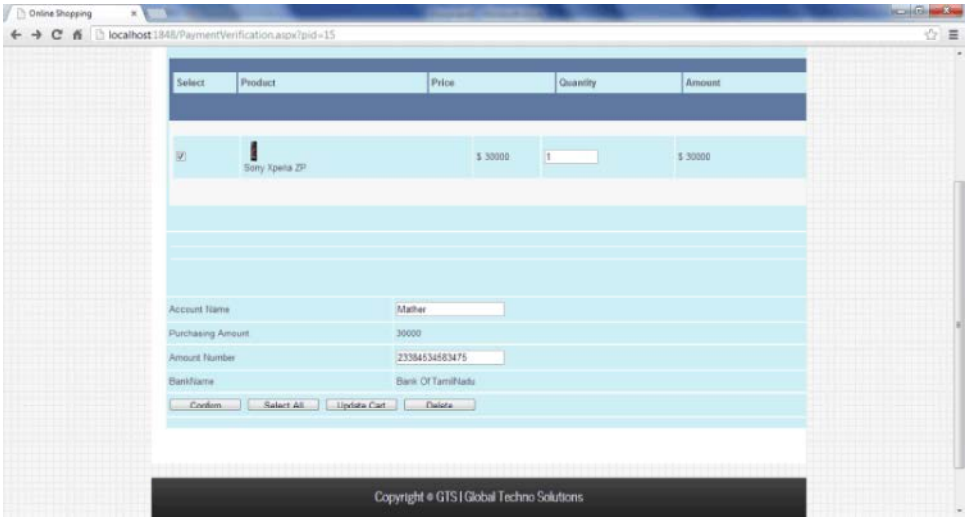


Fig. 6: Cart Details

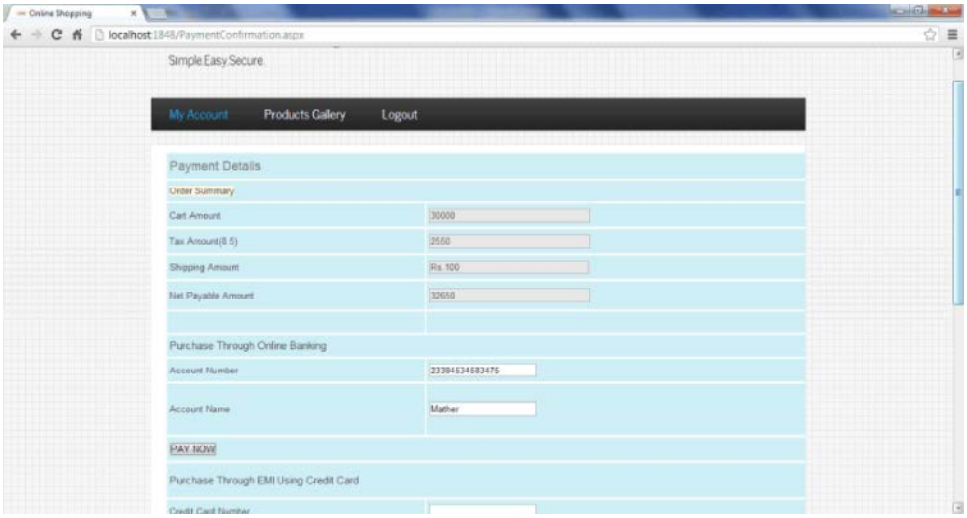


Fig. 7: Payment Details

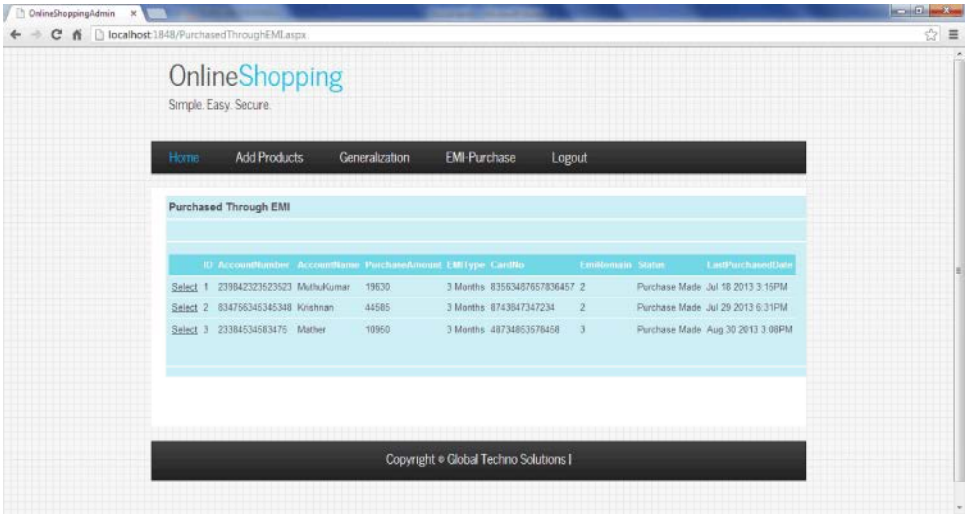


Fig. 8: Emi Purchase Request

The Figure5 shows transaction through the date. By giving date can get transaction of the particular date.

The Figure6 shows cart details. After selection of the product the user add the product in to cart in order to purchase. Then users have to give their account details.

The Figure7 shows payment details. After purchasing the product to make online payment the payment details need to fill by user.

The Figure8 shows EMI purchase request. The users having EMI purchase option also. The admin can view the EMI purchase details.

CONCLUSION

In this paper two phase top down method is proposed to give ability to handle large amount of data sets. By using effectual anonymization methods provide the privacy to cloud. Privacy preserving data publishing and data analysis are becoming severe issues in today's world. So the various techniques of data anonymization approaches are proposed. TDS method of Map Reduce applied into cloud for data anonymization along with deliberately planed group of inventive Map Reduce works to concretely achieve the specialization calculation in highly scalable. TDS method is high efficient and scalable.

REFERENCES

1. Chaudhuri, S., 2012. 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.
2. Zhang, X., C. Liu, S. Nepal, S. Pandey and J. Chen, 2012. A Privacy Leakage 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.
3. Mohan, P., A. Thakurta, E. Shi, D. Song and D. Culler, 2012. Gupt: Privacy Preserving Data Analysis Made Easy, Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD'12), pp: 349-360.
4. Mohan, P., A. Thakurta, E. Shi, D. Song and D. Culler, 2012. Gupt: Privacy Preserving Data Analysis Made Easy, Proc. ACMSIGMOD Int'l Conf. Management of Data (SIGMOD '12), pp: 349-360.
5. Hsiao-Ying, L. and W.G. Tzeng, 2012. A Secure Erasure Code-Based Cloud Storage System with Secure Data Forwarding, IEEE Trans. Parallel and Distributed Systems, 23(6): 995-1003.
6. Zissis, D. and D. Lekkas, 2011. Addressing Cloud Computing Security Issues, Future Generation Computer Systems, 28(3): 583-592.
7. Roy, I., S.T.V. Setty, A. Kilzer, V. Shmatikov and E. Witchel, 2010. Airavat: Security and Privacy for Mapreduce, Proc. Seventh USENIX Conf. Networked Systems Design and Implementation (NSDI'10), pp: 297-312.
8. Takabi, H., J.B.D. Joshi and G. Ahn, 2010. Security and Privacy Challenges in Cloud Computing Environments, IEEE Security and Privacy, 8(6): 24-31.
9. Fung, B.C.M., K. Wang, R. Chen and P.S. Yu, 2010. Privacy-Preserving Data Publishing: A Survey of Recent Devel- opments, ACM Computing Surveys, 42(4): 1-53.
10. Jurczyk, P. and L. Xiong, 2009. Distributed Anonymization: Achieving Privacy for Both Data Subjects and Data Providers, Proc. 23rd Ann, IFIP WG 11.3 Working Conf. Data and Applications Security XXIII (DBSec '09), pp: 191-207.
11. LeFevre, K., D.J. DeWitt and R. Ramakrishna, 2008. Workload-Aware Anonymization Techniques for Large-Scale Data Sets, ACM Trans. Database Systems, 33(3): 1-47.
12. Iwuchukwu, T. and J.F. Naughton, 2007. K-Anonymization as Spatial Indexing: Toward Scalable and Incremental Anonymization, Proc. 33rd Int'l Conf. Very Large Data Bases (VLDB '07), pp: 746-757.
13. Fung, B.C.M., K. Wang and P.S. Yu, 2007. Anonym zings Classification Data for Privacy Preservation, IEEE Trans. Knowledge and Data Eng., 19(5): 711-725.
14. Jiang, W. and C. Clifton, 2006. A Secure Distributed Framework for Achieving k-Anonymity, VLDB J., 15(4): 316-333.
15. Microsoft Health Vault, http:// www.microsoft.com/ healt/ ww/ product/Pages/healthvault.