I J C T A, 9(7), 2016, pp. 3095-3102 ISSN: 0974-5572

Research Paper on Privacy Preservation by Data Anonymization in Public Cloud for Hospital Management on Big data

P. Shyja Rose*, J. Visumathi and H. Haripriya*****

ABSTRACT

Big data has brought a revolution in the universe of data conclusive. If the data is shown to the other side people means, it violates the privacy of the information. So it is need to provide the privacy over the personal sensitive information's. Here we consider the dataset as electronic health care records, which contain the personal sensitive information's. Most of the existing systems are failed because of scalability, utilization of data and security of data on the public cloud. In proposed system if the information is transfer from one place to another means how to securely transfer the data and also how to provide the privacy over the sensitive information's. For administration, they do not need to build their own up infrastructure. It can reduce the cost of consumption. All the data's are stored on the cloud in the encrypted form by using an efficient encryption algorithm. Hybrid bunching technique is introduced for overcome the problem of existing system. To gain high scalability of data MapReduce algorithm is used. For keeping both data confidentiality and patient's identity on public cloud a organization novel authorized accessible privacy model is used, which can provide the specific access of data on public cloud by setting an access tree.

Key words: Big data, Sensitive information, Encryption, MapReduce, public cloud, Hybrid clustering, AAPM.

1. INTRODUCTION

Many organizations now using daily updatable or changeable data. For keeping data both security and usability cloud computing provide the environment to store data on different cluster.. Various organizations (e.g., Hospital authorities, industries and government organizations etc) freeing person specific data, which called as private sensitive information. They provide information of privacy of persons. The value hidden in big data can be of great value to cyberpunks and invaders. So, there is a gap between big data availability and big data security.

The main need of preserving privacy is protecting individual's sensitive information on a public platform. Unluckily de-identification of persons even by neglecting denotative identity like name, SSN, Voter Id number and license number. Data anonymization is the best way to preserve privacy over the personal privacy sensitive information. This data anonymization approach is very efficient technique but if the scalability of the data set like private sensitive information is increased the anonymization technique fails to preserve privacy. It is very need to preserve privacy for Big Data since the property of 3V i.e Volume velocity and Variety. So we have to provide an environment of secure patient's details in cloud.

So we have to provide scalable big data privacy preservation in cloud. A patient gives a hospital permission to continue a medical history with the expected value that it will help medical care. If this very sore and personage data is passed over to an appears in the newspapers, the patient has bewildered control over a crucial look of his or her life.

* Student, Jeppiaar Engineering College, *Email: shyjarose12@gmail.com*

** Professor, Jeppiaar Engineering College, *Email: jsvisu@gmail.com*

*** Student, Jeppiaar Engineering College, *Email: hpriya026@gmail.com*

The purpose of this project is to develop an environment to securely transfer the electronic health care records and provide privacy over the personable sensitative data. The Major aim of the work is to develop a tool for patients to give medical care providers more insight into your personal health information. Main aim of privacy is the secure data at the same time provide external knowledge also. Usability of information is more important than the privacy. It is a tool that you can use to gather, chase patient's information's. This application also helps to view the patient's health records only for authorized persons.

Typically person-specific data, which is named micro-data, is stored in a table that each row (tuple) corresponds to one individual. Generally this table has 4 kinds of attributes:

- Assigns like person's name, SSN, Voter Id number and license number are identify persons very easily so this type of attributes are called as identity attributes.
- Assigns like designation details in an organization, salary details in an company and disease name in an hospital are important for individual and this kinds of attributes are called as sensitive attributes.
- Assigns like pin code, nationality, gender, age are exist in another databases and we can easily identify the persons information by combining this kinds of attributes are called quasi attributes.
- Other than the above three attributes which is important but it won't violate the patients privacy so this kinds of attributes are called normal attributes.

So it is need to maintain both security and privacy over the rest of data and transferred data on the outsourced database. A novel authorized accessible privacy model is used to provide the limited access of data to the user over an public cloud. To achieve scalability all algorithm is designed with MapReduce. A new approach called hybrid anonymization is used to achieve best privacy over the Big data.

2. LITERATURE SURVEY

Xuyun Zhang et al. [1] proposes providing security and privacy over the intermediate data sets become dispute problem since adversaries may retain micro data by identifying multiple data records. Encryption of all datasets in the public platform called cloud take in past techniques may very time consuming and costly. So we provide new novel upper bound privacy leakage constraint-based techniques to give which intermediate data records demand to be ciphered and which do not to insure some data utilization and privacy preservation. A graphical representation like tree is drawn from generation relationships of midterm datasets to check privacy propagation of individual data records. And then this trouble is spited into many sub problems by decomposition privacy leakage constraints. And at the last to identify the needed encrypted dataset is done by the practical heuristic algorithm.

Mohammad Reza Zare Mirakabad et al. [2] aims providing privacy over the data publication. Under privacy data utilization and prevention of disclosure of individual identity is more important. One of the data anonymization techniques called K-anonymity prevents the disclosure of individual identity but it is mostly failed to attain. The other technique called l-diversity will provide the security of sensitive information. So author [2] proposes That L-Diversity is Lonely Enough for Preserving Privacy. In this technique the anonymization is performed by assigning individuals in the group of size greater than or equal to the value of quasi identifier k.

Min Wu et al. [3] proposes preserving privacy is most essential but the same time it is trouble in release of micro data release. In the view of attribute disclosure K-anonymity is not well. So we propose new technique called an ordinal distance based sensitivity mind full diversity metric model. The diversity of sensitive attributes is done only on the K-anonymized table. To check the diversity degree of the cluster first we have to classify the attributes with respect to first, second and third level. And the diversity degree is equal to the entire table. This method is mainly focus on the categorical values.

Yunli Wang et al.[4] proposes k-anonymity fails to achieve attributes disclosure but in l-diversity aims to achieve attribute disclosure. Second data anonymization technique concentrate on cutting the illation from freed micro attributes. Here we aim a new approach called an unique distinct l-SR diversity to achieve l-diversity on the sensitivity degrees of sensitive attributes. This results show that our algorithm achieved better performance on reducing illation of sensitive information and achieved the comparable generalization data quality compared with other data publishing algorithms. Finally we have two measures i.e Entropy Metric and Variance Metric to check the quality of published data.

Jordi Soria Comas et al. [5] aims data anonymization techniques preserve privacy, k-anonymity and €-differential privacy are two main privacy model. Author [5] t-closeness is the extension of k-anonymity, the construction of private sensitive information is based on Bucketization algorithm. The main aim of t-closeness is to solve the assigns disclosure issues. A data records is said to complete t-closeness if, for each crowd of data dealing a joining of quasi-identifier attribute evaluate, the gap among the distribution of each confidential attribute in the crowd and the distribution of the same confidential assign in the all data records is no more than a limit t. The Bucketization construction is used to achieve t-closeness. If the algorithm of Bucketization is too harsh the data loss in the confidential assign is large.

3. SYSTEM DESCRIPTION

Architecture of the organization is as demonstrated in Figure 1. The data files released by the data owner that is patient is get by the data publisher and by using the basic data like name and mail id the data publisher can create their own unique id value. And also the information is shared with a unique id generated by the admin. The whole data records about the patient are split into multiple chunks whenever it exceeds the limit of 64MB. And then the chunk records are to be encrypted then stored by using the key values in different clusters.

The data files released by the data owner that is patient is get by the data publisher and by using the basic information's the data owner can create their own profile. Whenever the data records are needed by the data publisher they decrypted by algorithm of hashing. And also the information is shared with a unique id generated by the admin. The whole data records about the patient are split into multiple chunks whenever it exceeds the limit of 64MB.

And then the chunk records are to be encrypted then stored by using the key values in different clusters.

If the sharing of records is done between two hospitals along with the permission of existing hospital it is need to maintain the privateness of data for that the anonymization techniques used the data files after the anonymization techniques should maintained on the database. And it is very need to separate authorized person and indirectly authorized person and unauthorized person, depending upon the access tree method accessibility provide.

4. MODULE DESCRIPTION

4.1. Registration and Unique Identifier generation

This module is necessary for creating PHR tool. This is used to provide information's for all patients with a unique id values. All the employees in the hospital can fetch the data of patient only by the id from the database. This tool used to eliminate the cost of consumption also. For a hospital to maintain or create one patient medical records means we need a unique identifier value for the patient in an organization. For generating id value for each and every user the Md5 hash algorithm is used. That takes password as input and 32 bit id as an output.

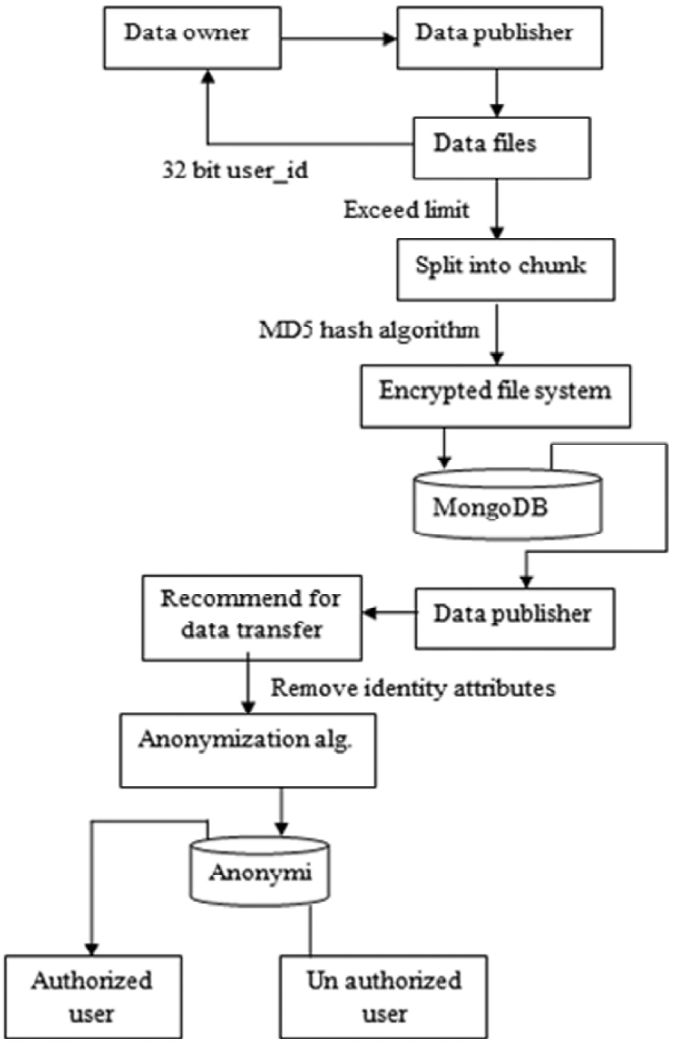

Figure 1: System Architecture

4.2. Data on Mongolab (public and private database)

Mongo DB application preparation is used to store patient's records. Each and every data along with this metadata are to be stored in a single place; it will simplifies the access time of data and minimize the use of joining the tough modules. User Files are automatically placed in the Public & Private Cloud (i.e. Mongo DB) based on the Sensitiveness of the Data. Normal Files are placed in the Public Cloud, Secured Data in Private Cloud. Multiple Mongo DB Cloud is spread for fast Data Retrieval.

4.3. MD5 hash algorithm–key values

MD5 processing 512-bit block as a input and produces a 128-bit (16 byte) blocks are often expressed as 32-digit hexadecimal value. After getting all the medical records from the patients it is need to store in an encrypted form at both rest and the transformation. All patient's sensitive medical information are change into key values by using an Md5 hash algorithm only for the limit. If the limits exceeds means splits into chunk and then perform hashing. After hashing the encrypted data are separately stored on different cluster.

4.4. Transfer medical records

The information of the patient's records is in encrypted form as normal so whenever we want to transfer medical records for patient's treatment we need to decrypt records in the opposite side. And also the first

hospital should allow the access of data files for easy access. For preserving privacy over the data the identity information's should be removed and the only need to perform anonymization techniques.

4.5. Hadoop MapReduce algorithm for anonymizing data

In our project we are dealing with massive amount of daily updatable patients records. For that Hadoop MapReduce algorithm is used. We are providing input (fuel), the engine converts the input into output rapidly and expeditiously, and you get the outputs you need. Hadoop MapReduce has several components like HDFS and MapReduce. HDFS is fully used for data storage it contain both data node and the name node. And the Map Reduce is used for performing the operation by Mapper and Reducer.

4.6. Anonymized patients records

In this module the anonymized patient record is generated by the technique of generalize and specification. In the technique of generalize the critical attribute of quasi domains value is changed into more general form. For example patient native place is Chennai means after the anonymization algorithm we get India as result. In suppression technique hide the values by symbols. For example pin code number has 6 digits changing values of number to symbol only for last three digits.

4.7. Authorized access of records

In this module by using a novel authorized accessible privacy model (AAPM) allows only the authorized person to view and edit the patient's records. The authorized person has full rights over the data. The indirectly authorized person has only rights to insert data rather than edit or view. But the unauthorized person has no rights over the data.

5. ALGORITHMS DEFINED

5.1. Md5 hash algorithm

MD5 processing 512-bit as stimulus value and generate 128 bit as production value are digest often expressed as 32-digit hexadecimal value. After getting all the medical records from the patients it is need to store in an encrypted form at both rest and the transformation. All patient's sensitive medical information are change into key values by using an Md5 hash algorithm only for the limit. If the limits exceeds means splits into chunk and then perform hashing. After hashing the encrypted data are separately stored on different cluster.

5.2. Hybrid data anonymization algorithm

Anonymization is the best way to provide privacy over the micro data which is released by the data publisher. Most of the existing systems are using several anonymization algorithms. Large number of top down specification and bottom up generalization techniques are developed separately for many organizations. We need to provide security and privacy over the patient's sensitive information. Categorical attributes are also available in the data sets for categorical attributes we introduce hybrid anonymization algorithm which include both TDS and BUG. Consider an example as follows figure 2.

Consider an example of 12 records with an attribute of education. And generate a taxonomy tree by using the attribute values.

Top down specification need to perform generalization for creating 4-anonymity table while bottom up generalization need not to do anything because the table of data records is already in 4-anonymity. At another end, let k be as 5 so the bottom up generalization need to perform generalization of changing child with the parent value and top down specification do nothing. The main aim of our

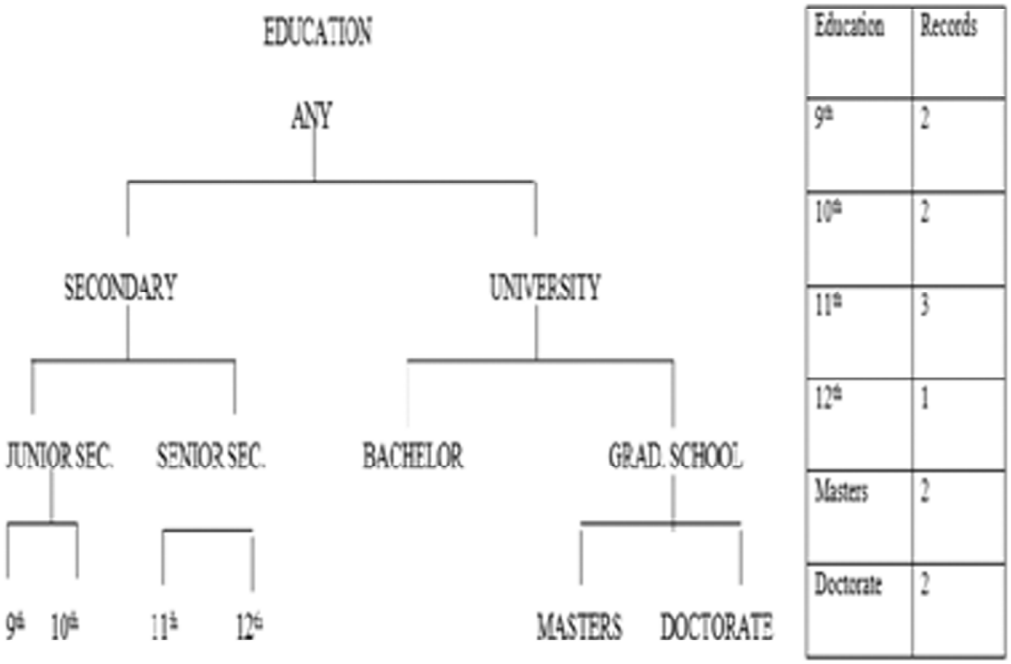

Figure 2: Taxonomy tree for Education

hybrid data anonymization is to create an environment of easy. Although a domain expert is able to understand which technique is most suitable to conduct anonymization manually according to the value of parameter k, ordinary users in the cloud likely break down to do this due to their lack of background knowledge.

$$IIPG(gen) = IL(gen)/(PG(gen) + 1)$$

$$IGPL(spec) = IG(spec)/(PL(spec) + 1)$$

Technique of generalize change the value of child with its parent in the taxonomy and in specification the value of parent is replaced with child values. Both techniques will be used only the trade off value is high. To find the trade off value the above *IIPG* and *IGPL* is to used.

6. RESULTS AND DISCUSSION

Over all performance of the project is finding out by comparing our PHR with existing systems. In an existing system Ucloud service provider is used. Instead of Ucloud in our PHR we are using Mongolab service provider for security. Mongolab encrypt and maintain replica of both transformation of data and rest of the data.

If the number of records is increased means allso service provider encrypts data records. Figure 3 shows the comparison and speed of Mongolab with Ucloud service provider. Both encryption and decryption of data also takes lot of time.

But in our Mongolab both encryption and decryption speed is minimum while compared with an existing system.In existing system uses two phase clustering algorithm for performing anonymization due to scalability and utilization it fails. For provide privacy and utilization we introduce hybrid anonymization algorithm which is nothing but a combination of both TDS and BUG algorithm and their performance also measured by some factors.

Figure 4 depicts the overall execution of our PHR system. Whenever we increase the number of records to store on the cloud also our system provides the minimum execution time by compare with the existing system. The privacy cost of our new hybrid anonymization algorithm is also low.

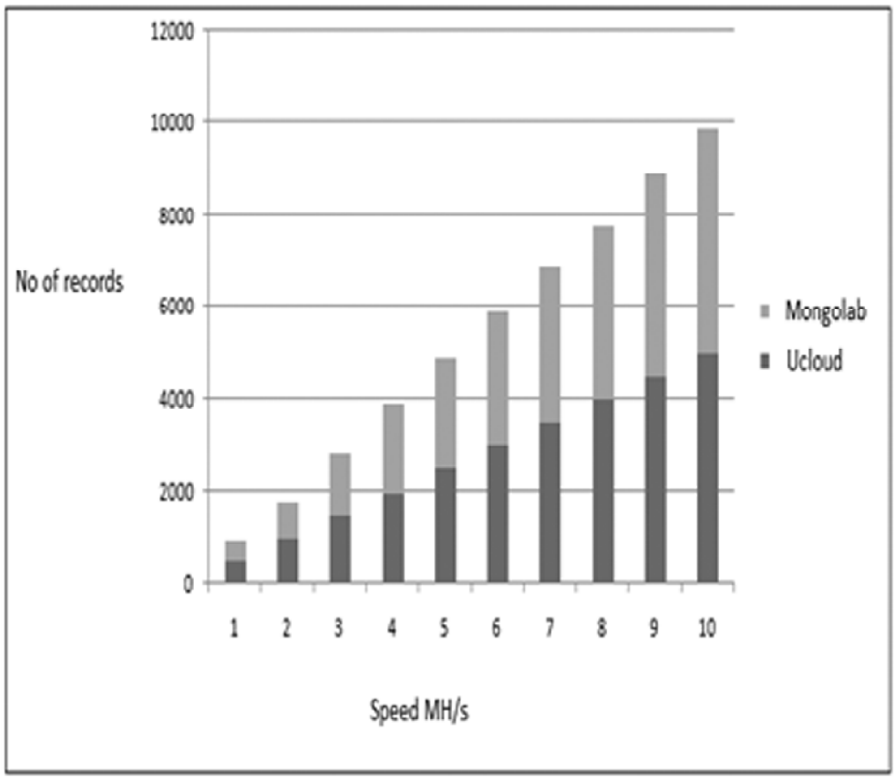

Figure 3: Performance of Mongo Lab Over U Cloud

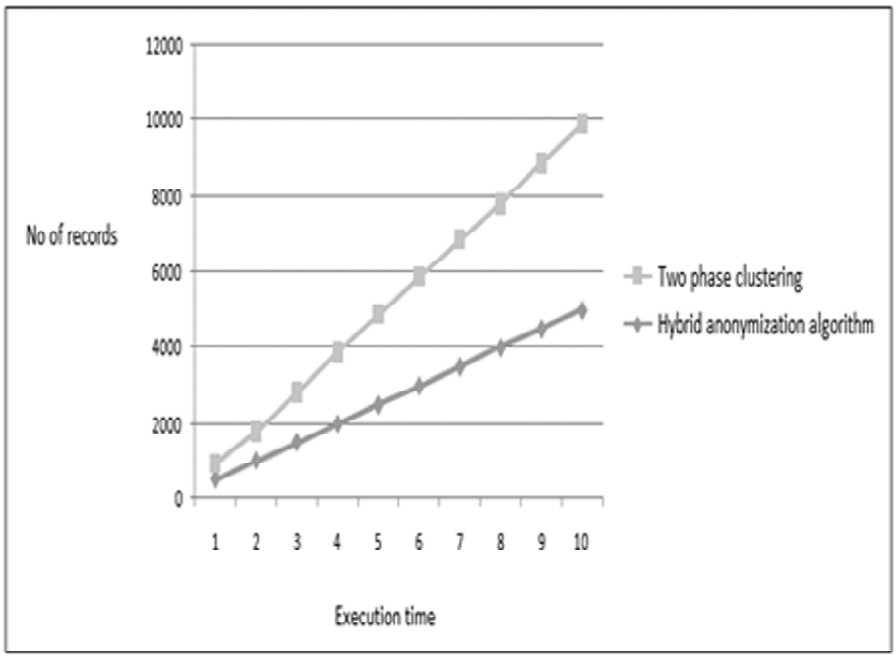

Figure 4: Overall execution time

7. CONCLUSION AND FUTURE WORK

There are many data anonymization approaches there for preserving privacy on micro data, each approach has its advantages and disadvantages. This survey paper has provided a comprehensive overview of privacy among the data which is placed on the public cloud. Finally the conclusion is the data anonymization techniques were failed because of the scalability and information loss. So we have to improve the scalability of the micro data. All the above mentioned algorithms are designed on the MapReduce to improve provide

the high scalable environment. Access of data from the public cloud can be secure only because of organization novel authorized accessible privacy model which can provide the access of data on public cloud by setting a access tree.

REFERENCES

[1] X. Zhang, C. Liu, S. Nepal, S. Pandey and J. Chen, "A Privacy Leakage Upper Bound Constraint-Based Approach for Cost-Effective Privacy Preserving of Intermediate Data Sets in Cloud," IEEE Transactions on Parallel and Distributed Systems, vol. 24, no. 6, pp. 1192-1202, 2013.

[2] Mohammad Reza Zare Mirakabad School of Computer Sciences, USM, Malaysia Intern at School of Computing, NUS, Singapore reza@cs.usm.my, reza.z@comp.nus.edu.sg "Diversity versus Anonymity for Privacy Preservation".

[3] Min Wu, Xiaojun Ye Institute of Information System and Engineering School of Software, Tsinghua University, Beijing, 100084, China "Towards the Diversity of Sensitive Attributes in k-Anonymity".

[4] Yunli Wang, Yan Cui, Liqiang Geng and Hongyu Liu "A New Perspective of Privacy Protection: Unique Distinct l-SR Diversity".

[5] Jordi Soria Comas and Josep Domingo Ferrer "Differential Privacy via t-Closeness in Data Publishing"

[6] Fangwei Luo, Jianmin Han, Jianfeng Lu and Hao Peng "ANGELMS: A Privacy Preserving Data Publishing Framework for Micro data with Multiple Sensitive Attributes".

[7] Gaoming Yang, Jingzhao Li, Shunxiang Zhang, Li Yu "An Enhanced l-Diversity Privacy Preservation".

[8] R. Magesh; T. Meyyappan "Anonymization Technique Through Record Elimination To Preserve Privacy Of Published Data".

[9] J. Li, R.C.W. Wong, A.W.C. Fu and J. Pei, "Anonymization by Local Recoding in Data with Attribute Hierarchical Taxonomies," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 9, pp. 1181-1194, 2008.

[10] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi and A.W.C. Fu, "Utility-Based AnonymizationUsing Local Recoding," Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data (KDD'06), pp. 785-790, 2006.

[11] Xuyun Zhang, Laurence T. Yang, Chang Liu, and Jinjun Chen "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud".

[12] Xuyun Zhang, Chang Liu, Surya Nepal, Chi Yang, Wanchun Dou, Jinjun Chen" Combining Top-Down and Bottom-Up: Scalable Sub-Tree anonymization over Big data using MapReduce on Cloud".

[13] Jiexing Li Yufei Tao Xiaokui Xiao "Preservation of Proximity Privacy in Publishing Numerical Sensitive Data".

[14] Robson L. F. Cordeiro, Caetano Traina Jr, Agma J. M. Traina, Julio López, U Kang, Christos Faloutsos. "Clustering Very Large Multi-dimensional Datasets with MapReduce".

[15] Max Grossman, Mauricio Breternitz, Vivek Sarkar "Hadoop CL: MapReduce on Distributed Heterogeneous Platforms Through Seamless Integration of Hadoop and OpenCL".

[16] Xuyun Zhang, Wanchun Dou, Jian Pei, Surya Nepal, Chi Yang, Chang Liu, and Jinjun Chen, Member, IEEE "Proximity-Aware Local-Recoding Anonymization with MapReduce for Scalable Big Data Privacy Preservation in Cloud".

[17] Jun Zhou, Xiaodong Lin, Senior Member, IEEE Xiaolei Dong, Zhenfu Cao, Senior Member, IEEE "PSMPA: Patient Self-controllable and Multi-level Privacy-preserving CooperativeAuthentication in Distributed m-Healthcare Cloud Computing System".