Article can be accessed online at http://www.publishingindia.com

Big Data Privacy Preservation Using Two Phase Top-Down Specialization Algorithm with Multidimensional Map Reduce Framework on Hadoop

Shalin Eliabeth S*, Sarju S**

Abstract

Big data privacy preservation is one of the most disturbed issues in current industry. Sometimes the data privacy problems never identified when input data is published on cloud environment. Data privacy preservation in hadoop deals in hiding and publishing input dataset to the distributed environment. In this paper investigate the problem of big data anonymization for privacy preservation from the perspectives of scalability and time factor etc. At present, many cloud applications with big data anonymization faces the same kind of problems. For recovering this kind of problems, here introduced a data anonymization algorithm called Two Phase Top-Down Specialization (TPTDS) algorithm that is implemented in hadoop. For the data anonymization-45,222 records of adult's information with 15 attribute values was taken as the input big data. With the help of multidimensional anonymization in map reduce framework, here implemented proposed Two-Phase Top-Down Specialization anonymization algorithm in hadoop and it will increases the efficiency on the big data processing system. By conducting experiment in both one dimensional and multidimensional map reduce framework with Two Phase Top-Down Specialization algorithm on hadoop, the better result shown in multidimensional anonymization on input adult dataset. Data sets is generalized in a top-down manner and the better result was shown in multidimensional map reduce framework by the better IGPL values generated by the algorithm. The anonymization was performed with specialization operation on taxonomy tree. The experiment shows that the solutions improves the IGPL values, anonymity parameter and decreases the execution time of big data privacy preservation by compared to the existing algorithm. This experimental result will leads to great application to the distributed environment.

Keywords: Big Data, Cloud Computing, Data Anonymization, Map Reduce, Privacy Preservation, Top Down Specialization

Introduction

This paper is a part of my project for the data anonymization in input adult dataset. The adult dataset was obtained based on a survey report. The personal information will not be hiding the sensitive information by lacks of proper anonymization algorithm in distributed environment. The sensitive input data can be preserved by using various algorithms on existing cryptographic mechanisms. Now day's researchers are trying to develop cryptographic algorithms by conducting experiments in hadoop. This data is properly studied and understand by analyzing the obtained information is publicly available or not. There are many anonymization algorithms and their aim is to protect the further identification of information on input dataset.

Big data is a large collection of information refers to data collection in applications have been growing tremendously and complicatedly. So that traditional data processing tools are incapable of handling the data processing

* Department of Computer Science and Engineering SJCET, Palai, Kerala, India. E-mail: shalinelizabeth9@gmail.com
** Department of Computer Science and EngineeringSJCET, Palai, Kerala, India. E-mail: sarju.s@sjcetpalai.ac.in

pipeline including collection, storage, processing, mining, sharing, etc. within a tolerable elapsed time. Nowadays, companies and organizations have been collecting huge amount of data containing various personal information via their products or services such as social network websites, online healthcare services and location-based services. There are three research challenges of privacy -preserving big data publishing in cloud computing, from perspectives of scalability, monetary cost and compatibility.

At present, the big data applications can be suitable for many industrial purposes with map reduce operations on big data. Accordingly, an increasing number of data mining or analytical tools and platforms are built on top of map reduce, e.g., scalable machine learning library Apache Mahout. However, none of traditional anonymization has been built on such a paradigm, while the published or shared data is usually consumed by big platforms or tools mentioned above. As a result, traditional anonymization approaches lack the compatibility to be integrated with the state-of-the-art big data mining or analytical tools in hadoop platform. The society will use sensitive information to distribute after anonymizing the number of customers using big data applications. This paper analyses on how to properly anonymize the big data by proper algorithm on multidimensional map reduce framework. A large set of anonymization algorithms was studied to overcome the problems of identifying sensitive identifiers on input adult dataset. The data set is 45,222 records of adult's information with 15 attribute values. But the network issues is the another problem in storing, processing and distributing the input adult dataset. By the analysis obtains that the smart home applications are the traditional way of privacy preservation for the identification purpose.

Relatted Work

In Big data applications, the privacy preservation for data analysis, share and processing is a challenging research issue due to increasingly large volumes of data sets, thereby requiring intensive investigation. A wide verity of privacy models and anonymization approaches have been put forth to preserve the privacy sensitive informational data sets. Data privacy is one of the most concerned issues because of processing large-scale privacy-sensitive data sets for big data applications.

Data Anonymization using One-Dimensional Map Reduce Framework

In this paper [1] the proposed data anonymization algorithm was implemented using one dimensional map reduce framework. If the dataset is so high, then the proposed anonymization algorithm does not work with the one dimensional map reduce operation. The cloud services on hadoop will requires users for different types of big data applications. Here introduces an anonymization algorithm called two-phase top-down specialization (TDS) algorithm for the privacy preservation by using single map reduce framework on distributed cloud environment implemented in hadoop. In map and reduce phases of operation, here uses single map reduce job for specialization computation with taxonomy tree by generating IGPL values.

Datafly Algorithm for the Data Anonymization

The Datafly algorithm [4] is one of the first practical applications in K-anonymity. The algorithm uses K-anonymity concept by generating the unique records of information in input dataset. The Datafly algorithm is performing with generalization and suppression steps to make data ready for release. The specialization operation is for providing K or less tuples on the data set with taxonomy tree. Step 3 is to suppress those tuples with frequency less than K. At last constructs the table of values based on K-anonymity parameter. The obtained domain is a hierarchy on taxonomy tree and the attributes of data records are generated by using specialization operation.

Mondrian Algorithms for the Data Anonymization

Mondrian [2] is for preserving privacy on data set with multidimensional map reduces approach that generalizes the data using sensitive attribute on IGPL approach. This approach is an efficient to anonymize a better approach than the other anonymizing schemes [3]. The Mondrian algorithm uses a specific anonymization mechanism by partitioning the input dataset by suppression methods. If we have the quasi-identifiers, then we need a dimension representation of Mondrian. There are two types of partitioning algorithms in Mondrian, i.e. the relaxed and the strict partitioning.

Methodology

Two-phase Top-Down Specialization (TPTDS)

There are 3 steps in the TPTDS approach, i.e.

(1) Data partition,

(2) Anonymization level merging

(3) Data specialization

Sketch of Two-Phase Top-Down Specialization

The TPTDS method is for privacy preservation which is required in TDS algorithm by obtains IGPL values. Generally map reduce on the cloud has two levels of parallelization i.e., job level and task level [4]. For example, the Amazon Elastic Map-Reduce service [5]. Task level parallelization is performed after the data partitioning operation [6]. A proper data partitioning algorithm is required for this. The second step is for obtaining the intermediate file by the TPTDS algorithm [7] and further anonymizes entire data sets. The subroutine is a map reduce edition of centralized TDS (MRTDS) [7] which concretely conducts the computation in TPTDS.

Data Partition

In the data partitioning algorithm [8], it is for obtaining values in the data records in *DI* and in *D*. In *m*-dimensional space, where *m* is the number of attributes obtained by anonymization algorithm. Random sampling technique is adapted to partition operation. Each file contains a random sample of *D*.

Anonymization Level Merging

All middle anonymization levels are merging into one in the second phase. The merging on anonymization levels [9] is completing by merging cuts. In multidimensional map reduce framework [10], here the intermediate results can merge by after obtaining information gain and the privacy loss.

Data Specialization

An original data set *D* [11] was obtained by the merged intermediate anonymization level *AL**. The IGPL Update job dominates the efficiency and the scalability of MRTDS [13], when it is executed iteratively as given in this method Thus, Hadoop variations [14] like Hadoop and Twist have been proposed recently to support efficient iterative map reduce computation.

Input: Data set *D*, anonymity parameters *k*, *k'* and the number of partitions *p*.

Output: Anonymous dataset *D**.

1. Performs the data partition on the input data *D* and obtains the intermediate dataset $Di,\ 1 \le i \le p$.
2. Execute MRTDS operation in parallel manner $(Di, k', AL) \rightarrow ALi,\ 1 \le i \le p$.
3. Performs the merging operation in obtained intermediate anonymization levels $(AL_1, AL_2, \ldots AL_p) \rightarrow AL^I$.
4. Execute MRTDS $(D, k, AL^I) \rightarrow AL^*$ to achieve k-anonymity by input anonymity parameter.
5. Performs the specialization on the dataset *D* according to *AL**, Output obtains as *D**

[1] The anonymization levels vary during in different stages of algorithm execution. Hadoop [15] using an open-source implementation of map reduce operation. The distributed cache execution in obtained data set is shown in Fig 1. MD5[17] (Message Digest Algorithm) is employed to compress the records transmitted for anonymity.

Map Reduce with Multidimensional Anonymization

Here propose a scalable map reduce based approach on multidimensional anonymization over big data sets [18] can see in figure 2. In this paper, introduces a scalable map reduce based approach for multidimensional anonymization over big data set [19] with seed initiation and seed updation algorithm. My basic and intuitive idea is to partition a large dataset recursively into smaller data partitions using map reduce until all partitions can fit in the memory of a computation node. Here multidimensional anonymization with three map reduce framework was proposed with seed initialization and seed updation algorithm [18]. But in this paper proposes the multidimensional anonymization with three map reduce frameworks on the MRTDS anonymization algorithm is shown in fig 2.

Fig 1: The Working of MRTDS Anonymization Algorithm on One Dimensional Map Reducing Framework

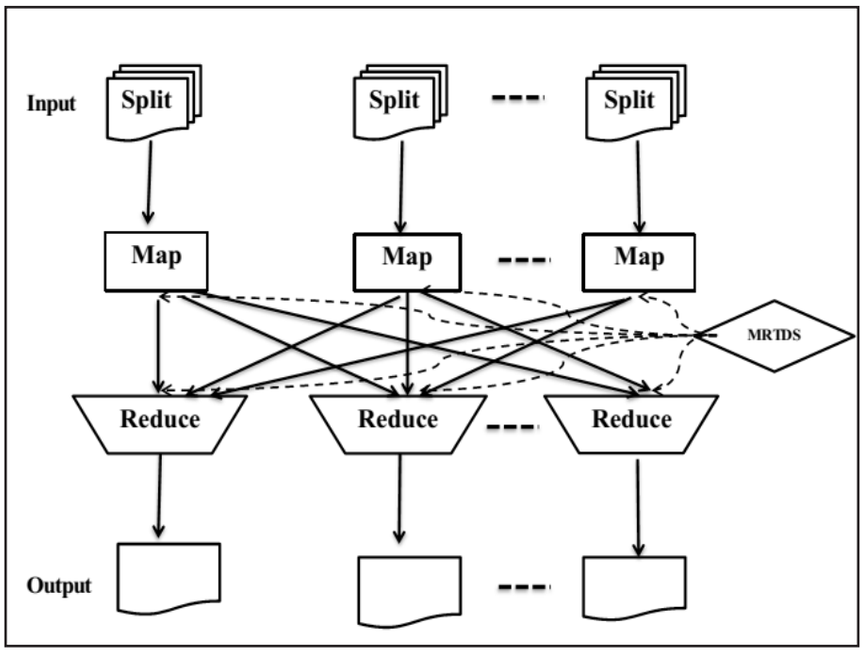
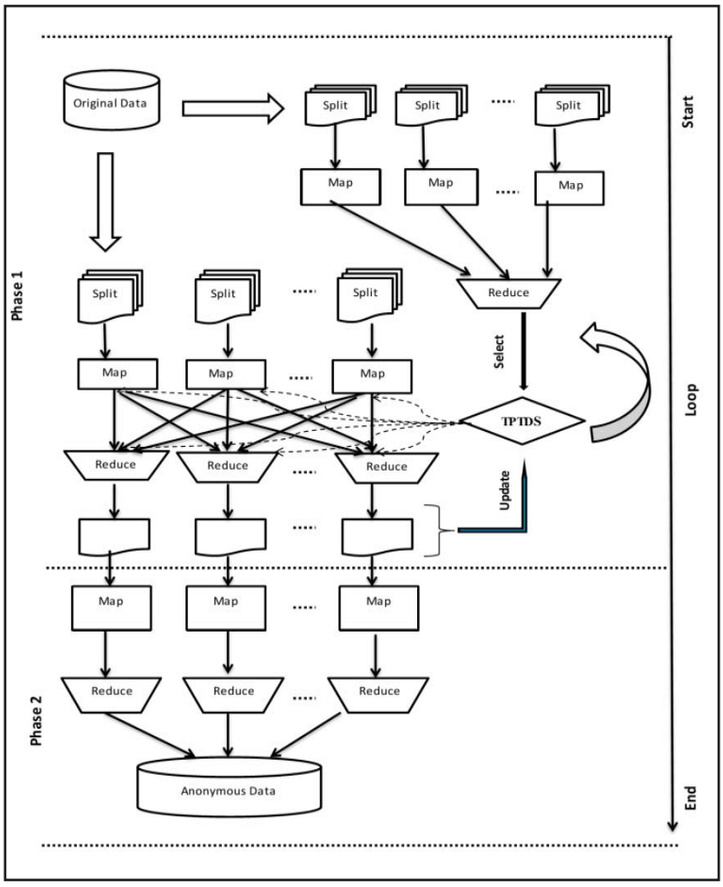


Fig 2: The Working of MRTDS Anonymization Algorithm on Multidimensional Map Reducing Framework.



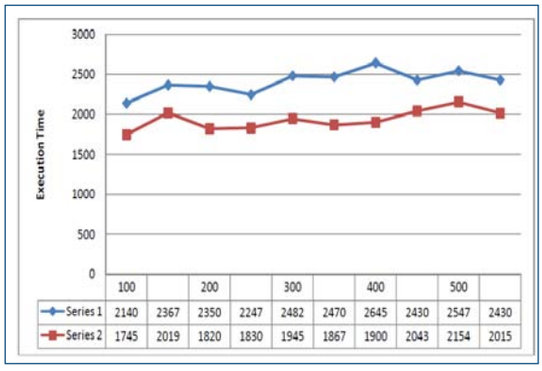
Results and Evaluation

In this section evaluates both the anonymization and information loss models. Here calculates the information gain and privacy loss by the TPTDS algorithm, unlike the existing papers of K-anonymity. In experiment shows that TPTDS multidimensional map reduce framework results are better in both case in information gain and privacy loss. Thus, all results presented in this paper is a result of the TPTDS multidimensional algorithm by comparing the one dimensional TPTDS anonymization. By comparing the base paper, the proposed TPTDS anonymization algorithm was implemented in multidimensional map reduce framework. But in case of base paper the TPTDS algorithm was implemented in one dimensional map reduce framework. The calculations are based on,

- Size of the dataset in MB
- Number of data partitioning
- Execution time
- Anonymity parameter k value
- IGPL values

These parameters are plotted in X and Y axis and compared the values in the base paper, by comparing the values in the graph, obtained the performance evaluation in one and multidimensional map reduce framework for the TPTDS anonymization.

Graph 1

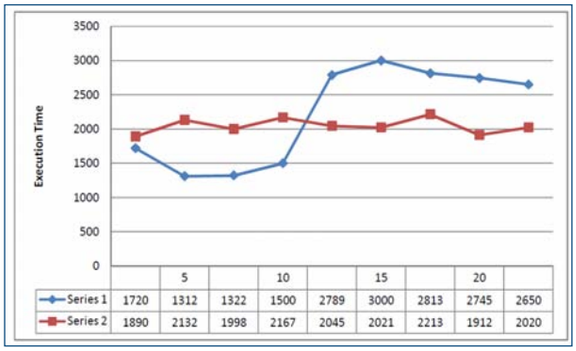


X-Axis: Size of the dataset in MB

Y-Axis: Execution time in seconds

The topmost line series 1 shown in the graph indicates that the TPTDS algorithm execution with one dimensional map reduce framework (In base paper). Bottom line series 2 indicates the TPTDS algorithm execution with multi dimensional map reduce framework (In proposed work). By comparing the two lines series in the 1st graph, the performance is better in multidimensional anonymization in series 2 by decreasing the execution time.

Graph 2

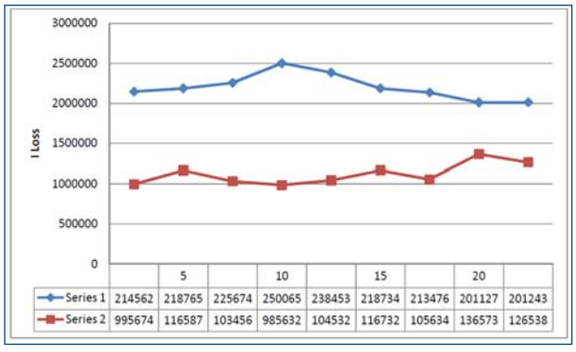


X-Axis: Execution time in seconds.

Y-Axis: Number of data partitions.

The line series 1 in graph 2 indicates that the TPTDS algorithm execution with one dimensional map reduce framework (In base paper). The line series 2 indicates the TPTDS algorithm execution with multidimensional map reduce framework (In proposed work). By comparing the two lines series in the 2nd graph, the performance is better in multidimensional anonymization in the line series 2 by decreasing the execution time.

Graph 3



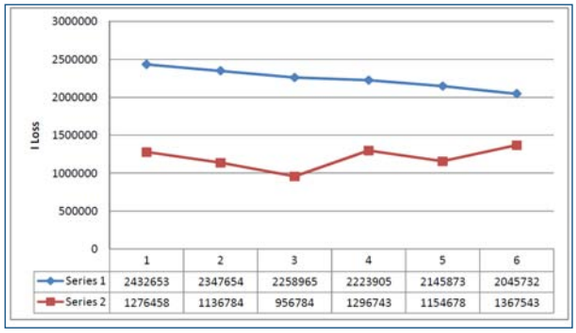
X-Axis: Number of data partitions.

Y-Axis: I Loss values obtained in TPTDS algorithm.

The series 1 line in graph 3 indicates that the TPTDS algorithm execution with one dimensional map reduce framework (In proposed work). The series 2 line indicates the TPTDS algorithm execution with multidimensional map reduce framework (In base paper). By comparing

the two lines in the 3rd graph, the performance is better in multidimensional anonymization in the series 2 line by decreasing the I Loss.

Graph 4



X-Axis: K-Anonymity parameter.

Y-Axis: I Loss in TPTDS algorithm.

The line series 1 indicates that the TPTDS algorithm execution with one dimensional map reduce framework (In base paper). The line series 2 indicates the TPTDS algorithm execution with multidimensional map reduce framework (In proposed work). By comparing the two lines in the 4th graph, the performance is better in multidimensional anonymization shown in the series 2 line by decreasing the I Loss. If the IGPL value increases then ILoss will also increases then degrades the performance on data anonymization.

Advantage on the proposed Multidimentional anonimyzation framework

1. Multidimensional scheme that recursively partitions the domain space to improve flexibility with regard to one-dimensional anonymization.
2. Improve the scalability and efficiency by indexing anonymous large data records by comparing the traditional methods.
3. It improves in time efficiency and which are cost effective in execution.
4. A method to reduce the required memory footprint.
5. Accurate in scheduling than the traditional map reduce framework.

Disadvantages

1. This approach is only effective in systems with high throughput, low-latency.
2. If the dataset is so high then the algorithm does not provide the better result.
3. High data splitting cause transmission overhead.
4. User constraints such as deadlines are important requirements which are not considered.

CONCLUSION

In order to overcome the existing problems, it is essential that a solution is devised. For existing one dimensional anonymization technique, the proposed multidimensional anonymization technique shows the better result. For providing security on the large dataset, here introduces Two Phase Top-Down Specialization anonymization approach using multidimensional map reduce framework on hadoop. The performance of the system is also evaluated based on the execution time. The experimental result is that the better result shows in anonymization with multidimensional map reduce framework by comparing the one dimensional map reduce framework. Here the map reduce operations was implemented in hadoop with multidimensional map reduce framework.

FUTURE WORK

There are several conditions that could be studied for the future. Risk evaluation could give an insight into how much information adversaries could dig out. In future it is expected to design the TPTDS anonymization by the bottom up generalization scheme on taxonomy tree for the data anonymization. And also the big data applications are trying to develop in twister. Many cloud computing vendors like Cloud era is offering HaaS (Hadoop as a Service). This can ease the configuration labour and can provide on-click elasticity of computational resources too.

REFERENCES

Zhang, X., Yang, L. T., Liu, C., & Chen, J. (2014). *A scalable two-phase top-down specialization approach for data anonymization using map reduce on cloud*. IEEE Transactions on Parallel and Distributed Systems (TPDS), 25(2), 263-373.

Zhang, X., Liu, C., Nepal, S., Pandey, S., & Chen, J. (2012). *A privacy leakage upper-bound constraint based approach for cost-effective privacy preserving of intermediate data sets in cloud*. IEEE Transaction on Parallel and Distributed Systems.

Zhang, X., Liu, C., Nepal, S., Dou, W., & Chen, J. (2012). Privacy-preserving Layer over Map Reduce on Cloud and Green Computing (CGC 2012), pp. 304-310, Xiangtan, China.

Jurczyk, P., & Xiong, L. (2009). *Distributed anonymization: achieving privacy for both data subjects and data providers*. Proceedings of 23rd Annual IFIP WG 11.3 Working Conference Data and Applications Security XXIII (DBSec '09), (pp. 191-207).

Liu H, Orban D (2011) *Cloud map reduce: A Map Reduce implementation on top of a cloud operating system*. In IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, (pp. 464-474).

Candan, K. S., Kim, J. W., Nagarkar, P., Nagendra, M., & Yu, R. (2010). *RanKloud: Scalable multimedia data processing in server clusters*. IEEE MultiMed, 18(1), 64-77.

Dean, J., Ghemawat, D. S. (2008). *Map Reduce: Simplified data processing on large clusters*. Communication of the ACM, 51, 107-113.

Fung, B. C. M., Wang, K., & Yu, P. S. (2007). *Anonymizing classification data for privacy preservation*. IEEE Transaction of Knowledge Data Engineering, 19(5), 711-725.

Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., & Fu, A. W. (2006). *Utility based anonymization using local recoding*. In ACM SIGKDD.

Jiang, W., & Clifton, C. (2006). A secure distributed framework for achieving k-anonymity. *VLDB Journal*, 15(4), 316-333.

Amazon Web Services. (2013). Amazon Elastic Mapreduce. Retrieved from http://aws.amazon.com/elasticmapreduce/ (accessed on January 05, 2013)

Roy, I., Setty, S. T. V., Kilzer, A., Shmatikov, V., & Witchel, E. (2010). *Airavat: Security and privacy for mapreduce*. Proceedings of 7th USENIX Conference on Networked Systems Design and Implementation (NSDI'10), (pp. 297-312).

Brodsky, A., Farkas, C., & Jajodia, S. (2000). *Secure databases: Constraints, inference channels, and monitoring disclosures*. IEEE Transactions on Knowledge and Data Engineering. 12, 900-919.

Cao, N., Wang, C., Li, M., Ren, K., & Lou, W. (2011). *Privacy preserving multi-keyword ranked search over encrypted cloud data*. Proceedings of IEEE Infocom, (pp. 829-837).

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., & Zaharia, M. (2010). *A view of cloud computing*. Communication of the ACM, 53(4), 50-58.

Mohan, P., Thakurta, A., Shi, E., Song, D. & Culler, D. (2012). *Gupt: Privacy preserving data analysis made easy*. Proceedings of ACMSIGMOD International Conference on Management of Data (pp. 349-360).

Hsiao-Ying, L., & Tzeng, W. G. (2012). *A secure erasure code-based cloud storage system with secure data forwarding*. IEEE Transactions and Distributed Systems, 23(6), 995-1003.

Zhang, X., & Dou, W. (2014). *Proximity-aware local-recoding anonymization with map reduce for scalable big data privacy preservation in cloud*. IEEE Transactions on Computers.

UCI Machine Learning Repository. Retrieved from ftp://ftp.ics.uci.edu/pub/machine-learnng-databases/