ANONYMIZATION OF DATA USING MAPREDUCE ON CLOUD

Mallappa Gurav[1], N. V. Karekar[2], Manjunath Suryavanshi[3]

[1]Dept. Of Computer Science and Engineering, K. L. E College of Engineering & Technology, Chikodi-591 201
[2]Dept. Of Computer Science and Engineering, K.L.S Gogte Institute Of Technology, Belagavi-590008, Karnataka, India
[3]Dept. Of Computer Science and Engineering, K. L. E College of Engineering & Technology, Chikodi-591 201

Abstract

In computer world cloud services are provided by the service providers. The user wants to share the private data which are stored in cloud server for different reasons like data mining, data analysis etc. These can bring the privacy concern. Privacy preservation can be satisfied by Anonymizing data sets through generalization to satisfy privacy requirements by using k-anonymity technique which is a widely used type of privacy preserving techniques. At present days the data of cloud applications are increasing their scale day by day concern with Big Data trend. So it is very difficult thing to accept, manage, maintain and process the large scaled data with-in the required time stamps. Thus for privacy preserving on privacy sensitive , large scaled data is very difficult task for existing anonymization techniques because they will not manage the scaled data sets. This approach addresses the anonymization problem on large scale cloud data sets using two phase top down specialization approach and MapReduce framework. Innovative MapReduce jobs are carefully designed in both phases of this technique to achieve specialization computation on scalable data sets. Scalability and efficiency of Top Down Specialization (TDS) is significantly increased over the existing approach.

Keywords: *Top Down Specialization, MapReduce, Data Anonymization, Cloud Computing, Privacy Preservation*

--***--

1. INTRODUCTION

In this work highly scalable "Two Phase Top Down Specialization approach" is used for anonymization of data based on MapReduce on cloud. Specialization is required in an anonymization process to make effective use of parallel computing capability of MapReduce on cloud. This process is split into two phases. In 1st phase the original data set is partitioned into different groups of smaller data sets and these smaller data sets are anonymized in parallel by the help of MapReduce to producing intermediate results. In 2nd phase the intermediate results which are generated in first phase are integrated into one data set, and this data set is further anonymized to achieve consistent k-anonymous data set. This approach tends MapReduce to achieve the concrete computation in both the phases. To do specializations on cloud data sets, a group of MapReduce jobs are carefully constructed and coordinated.

Cloud computing is a disruptive and rapidly growing trend at present status, which constitute a significant impact on present research communities and IT industry [1]. Cloud computing is giving huge computation power and large storage capacity through utilizing a more number of computer systems together and enabling the people to deploy applications cost effectively. The users need not to invest heavy infrastructure for small computations. Earlier the people are investing the large cost for heavy computation for some times, so it is not cost effective. Thus the cloud providers maintains the large infrastructure the users can use that infrastructure as they want (pay as you grow) without investing money and effort on setup of infrastructure. However many people are not taking the advantage of cloud because of privacy and security reasons [4]. The research has taking place now on cloud security and privacy has come to the picture [6]. In cloud computing the Privacy is one of the most factors. This issue is concern aggravates in the cloud computing environment and some privacy concern issues are not new [5]. The financial transaction record and health record of electronic are very sensitive for people. Humans can get profits from the sensitive data if they are analyzed and mined properly by some organizations like disease research labs. For example Microsoft HealthVault [10], in this they will take the data and aggregates it and shares with the research centers. Data privacy can be disturbed with very less efforts by cloud users in old privacy protection systems [5]. This can tends to damage the economics and social values of persons. Hence it is important to solve the privacy issue of data urgently in cloud before they are shared on cloud.

Anonymization of data is a technique has been deeply studied and widely adopted because of data privacy preservation in non interactive data publishing, sharing scenarios [11]. Data anonymization is a technique in which hides the identity or sensitive data of owner's data record. The user can use the data for many reasons like diverse analysis and mining then there is a need of privacy preservation of cloud data sets. Many anonymization algorithms (technique) with different anonymization operations have been proposed [12]. Day by day data sets scales are increasing rapidly in concern with Big Data and cloud computing that needs anonymizing in some cloud applications [1]. Traditional algorithms for anonymization are difficult task for anonymizing the Data sets in clouds. The research people are working to fix up the scalability problem of large scale data anonymization.

Many applications need powerful computation capability. For large scale data processing there is a need of frame work, which is provided by MapReduce and is integrated into cloud computing. So, it is important to adopt like frameworks to address the scalability problem of anonymizing large-scale data for privacy preservation [9] in cloud. To address this type of scalability problem of large scale data anonymization the MapReduce technique is used. The TDS approach, gives a good tradeoff in between data consistency and data utility which is widely applied for data anonymization [10]. Now a day the TDS (Top Down Specialization) algorithms available are centralized one and they are not properly handing the large scaled data sets. Some distributed algorithms are handling the secure anonymization but they are not concentrated on scalability aspect. As the MapReduce computation framework is relatively simple but it is very difficult to design MapReduce jobs.

In this methodology it is having threefold. First, creatively applying the MapReduce jobs on large scale data to anonymization, for we are carefully creating the MapReduce jobs to achieve the specialization as TDS in highly scalable fashion. In Second phase, a two phase TDS technique is applied to obtain the high scalability which is allowing specialization conducted on different data partitions which is created at first phase in parallel fashion. In third stage getting the experimental results significantly and analyzing it with existing approaches.

2. RELATED WORKS

When a problem has occurred, the solution to the problem demands the information which is not always necessary to find the solution in hard way, i.e. to gather the data, empirically. The first view could be to see that someone had worked on the same problem and got a good result on it. The second view that someone documented the solution in article or book, then there is a way to find that document to our work by reference search. Often this technique has remained as tactical knowledge of the trade or profession. The best case would be to study the relevant literature that can reveal some solutions. To get the solution we do not need to resort the empirical work. The study of literature is good thing that save human effort and money. Comparing the empirical work with literature study, the literature study is simple, cheaper and faster than empirical work. Therefore, most research works start with a study of literature.

Literature Survey for this project includes following study:

2.1 Scalability Problem of Anonymization Algorithms

The data of cloud is tremendously increasing in its size, for anonymization of these data is difficult. The author LeFevre et al. [7] introduced the “scalable decision tree” and “sampling technique”, these are used to addressed the scalability problem of anonymization algorithms. One more author Iwuchukwu and Naughton [8] introduced “R-tree index-based approach” by building a spatial index over data sets. This approach used for achieving high efficiency. However, the above mentioned approach aims at multidimensional generalization [5], but they are failing to work in the TDS approach. TDS approach is the technique in which the data are processed from top to down of architecture. Repetition of processing is applied on the data sets for anonymization.

2.2 TDS (Top Down Specialization)Approach

The author Fung et al. [2] presented one approach called as “TDS approach”. This approach produces the anonymous data sets without the data exploration problem [11]. A data structure “Taxonomy Indexed PartitionS (TIPS)” is exploited to improve the efficiency of TDS. But the approach is centralized, leading to its inadequacy in handling large-scale data sets. In centralized approach all the data sets are stored in central storage unit and processing can be occurred with this centralized data sets. By this the data retrieval can be slow and probably parallelism cannot be applied properly. Then we need the distributed processing system.

2.3 Distributed Algorithms

Many distributed algorithms are presented for privacy preservation of different data sets retained by different parties. For anonymizing the vertically partitioned data sets two authors Jiang and Clifton [4] and Mohammed et al. [2] proposed the distributed algorithms, where the data collected from different sources without disclosing the privacy of data from one party to other. For anonymizing the horizontally partitioned data sets two authors Jurczyk and Xiong and Mohammed et al. [2] given distributed algorithms retained by multiple holders. Hence the above mentioned distributed algorithms mainly focus on securely integrating and anonymizing the different data sources. This work mainly focuses on the scalability issue of TDS anonymization, and is, therefore, orthogonal and complementary to them.

2.4 Mapreduce Privacy Preserving

The author Roy et al. investigated the privacy problem associated with MapReduce in concern with privacy protection. Author presented the system “Airavat incorporating mandatory access control with differential privacy”. Another author Zhang et al. [7] worked on hybrid cloud; exploit MapReduce in which the computing jobs are partitioned into security levels of data. This work leverage the MapReduce for anonymization a large scale data values before they are processed by other MapReduce jobs which are arriving at privacy preservation of data sets.

By using above techniques we are working on efficient anonymization technique.

3. TWO PHASE TOP DOWN SPECIALIZATION APPROACH

The proposed Data Anonymization using MapReduce on Cloud contains five modules works in different stages. The modules are Data loading, Data Partition, Anonymization, Merging, and Specialization. This approach uses Two Phases for dealing with anonymization. In the first phase the collected large set of data is partitioned and each partitioned data set is applied by specialization. In second phase the intermediate generated data results are again applied by specialized technique and grouped together to form single result which is anonymized one.

The architectural overview of the project shows the flow of data or process between dataset with algorithm and MRTDS module, which is shown in Fig. no. 3.1. There are mainly five modules in this project namely data loading, data partition, anonymization, merging, specialization.

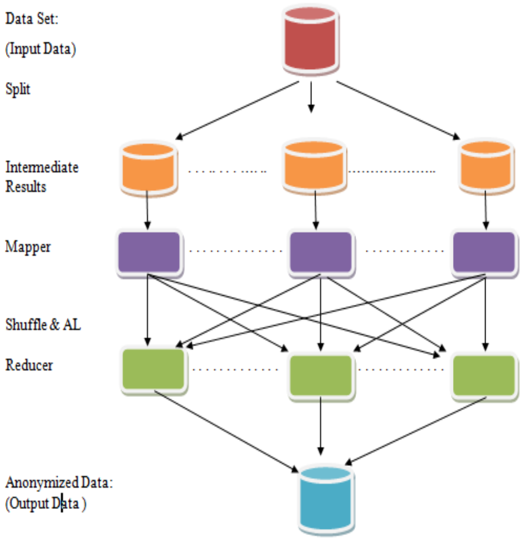


Fig 3.1 Architectural Overview

3.1 Load Dataset Module

The load dataset module should load the data set from the system. The user should select the data set which is stored in the predefined drive. The data is in predefined format, which having six attributes with it. When the user selects and loads the data sets is stored in the MySql data base.

3.2 Data Partition Module

Data partition module is required here, because the loaded data should be partitioned into heterogeneous datasets. Large no of data set is not working properly in MapReduce environment so we have to split the large data into smaller data sets. The splitted files which cannot have any common attribute values means intersection of splitted files is null. For generating the partition the random number is chosen for each data set. According to the random no we will store the data into different files.

3.3 Anonymization Module

Anonymization means hide or remove sensitive information from the data set for preserving the privacy. We are using k-anonymity methods for anonymizing. K-anonymity means the attribute values represents column values in datasets, the parameter k is received from the user for anonymization according to the anonymity parameter we can hide the sensitive information. In this module we first display the file content with it's attribute values and applying the three anonymization techniques on the data sets and displays the result. Anonymization can be applied on intermediate data sets also on merged data sets.

3.4 Merging Dataset Module

The partitioned data sets are merged here in this module. All the intermediate results generated are merged to form a single data set. The merging of anonymization level AL is done here. Merging operation is completed by merging the cuts.

3.5 Specialization Module

In this module the data record will be the input. The data record will be from data set D. The anonymization level is also the input for this module. By mapping the record we can find the Parent attribute. Finally the parental and children attribute values are replaced by anonymization parameter with original data sets by giving the anonymous data records.

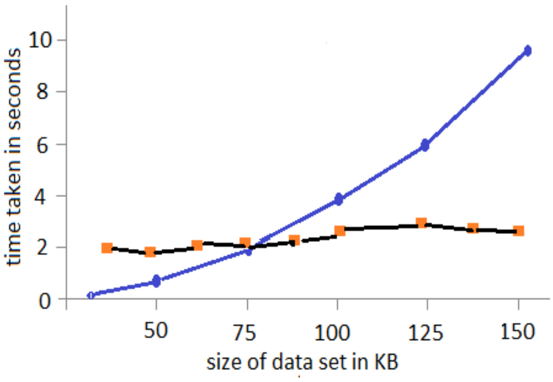
3.6 Sketch of Two-Phase Top-Down Specialization

The propose TPTDS algorithm is used computation required in TDS is highly scalable and efficient one. In above mentioned algorithm D is the original data sets which is inputted to the algorithm. K and k' are the anonymity parameters user should enter these, p is the no of partitions, AL is the anonymization level, D* is the anonymized data set as output.

```
ALGORITHM: SKETCH OF TWO-PHASE TDS (TPTDS).
Input: Data set D, anonymity parameters k, $k^{I}$ and the number of partitions p.
Output: Anonymous data set $D^{*.}$
1:  Partition D into $D_i$, 1 <=i <= p.
2:  Execute MRTDS(Di, k',$AL^0$) →$AL_i^1$ 1<=i<=p parallel as multiple MapReduce jobs.
3:  Merge all intermediate anonymization levels into one, merge(< $AL_1$', $AL_2$',……$AL_p$'>)→AL'
4:  Execute MRTDS(D, k, AL') → AL* to achieve k-anonymity. Specialize D according to $AL^*$, Output $D^*$.
```

This project is developed using the Java language framework. The system can be run from any platform. The OS chosen for the project is Microsoft Windows. Since Java is purely Object Oriented language and easy to implement any application. WAMP server software is used for database storage. Eclipse is used for better user interface for java programming language. Apache Hadoop for implementation of distributed storage and distributed processing.

4. RESULTS



R 🟧 ~~resents the I~~ 🟧 allel Execution

R 🟪 ~~resents the C~~ 🟪 ıtralized Execution

The above mentioned figure represents the results after conduction of repeated executions with different values. This experiment results shows that the size of data consider as x axis and time consumed for execution as y axis. The time consumed by parallel as well as centralized system is calculated. By the diagram it is clear that the time consumed for parallel execution is good concern with centralized approach. The centralized approach takes more time when the data size is increasing, but the parallel execution is not like that, it is consume approximately same time even the data size is increasing.

5. CONCLUSION AND FUTURE SCOPE

In this approach we have investigated the privacy issue in the cloud data. We proposed the system “Anonymization of data using MapReduce on cloud”. This system is working on highly scalable data sets using TPTDS approach. This system is having two phases in 1st phase the data sets are partitioned and anonymized into smaller, similar data sets parallel and producing intermediate results. In second phase all the intermediate results are combined and anonymized to produce k-anonymous data sets. In this work we have applied MapReduce on cloud data to anonymization and carefully designed the set of MapReduce jobs to achieve the specialization computation in highly scalable fashion. Set of experiments are conducted on real world data sets and the experimental results reveal that the anonymization is carried on data set effectively and scalability and efficiency of TDS are improved over existing approaches.

In cloud environment day by day the volume of the data is increasing, privacy preservation for analysis, share and mining of data is challenging research issue. So it needs the intensive investigation. In the future days this approach can be guidelines to develop bottom-up generalization algorithms for data anonymization. In this work we used MySQL database but in further days without this can be done. Based on the usage of guidelines provided from this system developer can work on scalable privacy preservation, 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 CloudComputing 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 Approach for Cost-Effective Privacy Preserving of Intermediate Data Sets in Cloud,”IEEE Trans. Parallel and Distributed Systems, to be published, 2012.

[7]. L. Hsiao-Ying and W.G. Tzeng, “A Secure Erasure Code-BasedCloud 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.

[9]. P. Mohan, A. Thakurta, E. Shi, D. Song, and D. Culler “Gupt: Privacy Preserving DataAnalysis Made Easy,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD’12), pp. 349- 360, 2012.

[10]. Microsoft Health Vault, http://www.microsoft.com/healt/ww/product/Pages/healthvault.

[11]. B.C.M. Fung, K. Wang, R. Chen, and P.S. Yu, “Privacy-Preserving Data Publishing: A Survey of Recent Devel- opments,” ACM Computing Surveys, vol. 42, no. 4, pp. 1-53, 2010.

[12]. B.C.M. Fung, K. Wang, and P.S. Yu, “Anonymizing Classification Data for Privacy Preservation,” IEEE Trans. Knowledge and Data Eng., vol. 19, no. 5, pp. 711-725, May 2007.

BIOGRAPHIES

Mr. Mallappa Gurav received the Bachelor of Engineering degree in Computer Science and Engineering from Basaveshwar Engineering College, Bagalkot, affiliated to VTU, Belgaum, during 2010. He is persuing M.Tech at Gogte Institute of Technology, Belgaum, under Visvesvaraya Technological University, Belgaum. Currently he is working as an assistant professor in Computer Science and Engineering Department of K. L. E. College of Engineering and Technology, Chikodi since from 2010.

Mr. N. V. Karekar, received his M.Tech degree in computer science and engineering and currently working as an assistant professor in Computer Science and Engineering Department of K. L. S. Gogte Institute of Technology, Belgaum, affiliated to Visvesvaraya Technological University, Belgaum. He had eight years of teaching experience and one year of industrial experience.

Mr. Manjunath Suryavanshi received a Bachelor of Engineering degree in Computer Science and Engineering from Gogte Institute of Technology, affiliated to VTU, Belgaum, during the year 2009. He completed his M.Tech. degree in Software Engineering from M. S. Ramaiah Institute of Technology, an autonomous institute affiliated to VTU, Belgaum, during the year 2013. He is working as an assistant professor in Computer Science and Engineering Department of K. L. E. College of Engineering and Technology, Chikodi, since from August-2011.