Privacy Preserving using data Anonymization technique on Hadoop

[1]Ms.Aarti Gopal Bhirud, [2] Prof.Harish Barapatre, [3]Prof.Nilima Nikam

[1.2.3.] Yadavrao Tasgaonkar Institute and technology Bhivpuri, Karjat

ABSTRACT

At present applications are developed widely and also increase various internet based services and cloud applications, the need of cloud environment with buildup facilities is increasing with very widely. Due to the increase in multiple users. A large number of cloud services require users to share private data like hospital health records for data scrutiny or mining, bringing privacy concerns. At present, the scale of data in many cloud applications increases hugely in accordance with the Big Data trend, thereby making it a challenge for commonly used software tools to manage, process, and capture such large-scale data within a tolerable elapsed time. It is challenging for previous annonymization approaches to acquire privacy on large scale data sets due to insufficiency.To preserve privacyamong the data being shared an adequate anonymization technique is used. In the proposed system we can apply data partition MapReduce method.Anonymization the data is important because in some cases,there is no way to stop this information from being available and some data has high value and analyzing it could lead to progress in fields that are highly important, like hospital and education related data. A scalable two-phase top-down specialization (TDS) approach uses MapReduce architecture on cloud to annonymized large scale datasets to design a group of innovative MapReduce jobs to particularly accomplish specialization computation in a highly scalable way. Anonymizing data sets via generalization to satisfy privacy requirements such as k anonymity will be use category of privacy preserving techniques. So the ability of TDS and efficiency of TDS can be significantly upgraded over existing approaches. We can solve scalability problem of large-scale data anonymization by TDS, and can proposed a two-phase TDS approach using MapReduce on cloud. In this paper we are focuses on providing authorized data and preserved identity of authorized data also protect the privacy ofindividuals represented in the data.. In TDS approach data sets can partitioned and anonymized in parallel in the first phase, and can generate midway result.Then,we can merged midway result and can further anonymized to yield consistent k-anonymous data sets in the second phase. Experimental evaluation results demonstrate that with our approach, the scalability and efficiency of TDS can be significantly improved over existing approaches

Keywords:-Big data ,Data anonymization,privacy preservation,Cloud,Mapreduce

1.NTRODUCTION

Internet is growing at a remarkable rate in recent years not just in terms of size but also in the term of provided.In the present day scenario poses a significant impact on current IT industry.The scale of data in many cloud applications increases enormously in accordance with the big data trend, thereby making it a challenge for commonly used software tools to capture, manage, and process such large-scale data within a tolerable elapsed time.Big data is a term given to a large and complex data that organizations store and process.However it is difficult for companies to store, retrieve and process the ever-increasing data.Every day we create large scale of data through various sites, also weather reports, purchase transaction records etc.Insted of depend on expensive hardware and different system to store and process data, hadoop enables distributed parallel processing of huge amount of data across various industry standard severs stores and process the data and can scale without limits.Hadoop can handle all type of data like structured and unstructured,images,email,audio files,log files etc.These data can contain information of individual for eg.Social security number and also contain personal information (like, date of birth,zip code,age,gender) that are probably identifying when link with other available data set.This paper defines the importat issue of preserving privacy by anonymity of the individuals or entries during the data distribution process.The focus of this work is to perform anonymization on data being shared on cloud that provide enough privacy on big data.Data anonymization refers to hiding identity and/or sensitivedata for owners of data records.The major focus of this system that all people could easily access the data shared on the public cloud and they can share their own data with the hadoop environment with confidently and the sensitive data they think that are not to be publish directly to the public can be protect from others by generalising those data.This can be implemented by specializing the level of information in a top-down manner until minimum privacy requirement is violated.Map-Reduce, a widely-adopted parallel data processing method, to address the scalability problem of the Top-Down Specialization (TDS) approach [7] for large-scale data anonymization. The TDS approach, offering a good adjustment between data utility and data consistency, is widelyapplied for data anonymization [7,8,12,13]. Most TDSalgorithms are centralized, resulting in their inadequacy in handling large-scale data sets.In this paper, we proposed a highly scalable two-phase TDS approach for data anonymization based on MapReduce on cloud. Top down Specialization required in an anonymization process are split into two phases. In the first one, original datasets are partitioned into a group of smaller datasets, and these datasets are anonymized in parallel, producing intermediate results. In the second one,the intermediate results are merged into one, and further anonymized to achieve consistent *k*-anonymous [09] data sets.

2.RELATED WORK AND PROBLEM ANALYSIS

Scalability problem of anonymization algorithms via scalable decision tree. An R-tree index based approach by creating spatial index over data set and achieving high efficiency. Therefore, the above approaches aim at multidimensional generalization[14],thereby failing to work in TDS approaches.Fung et a[7,8,12] proposed the TDS approach that produce anonymous data sets without the data analysis problem[10].A data structure Taxonomy indexed partitions is to improve the efficiency of TDS.But the approach is centralized,leading to insufficiency in handling large scale data set.

In the existing system we analyze the scalability problem of existing TDS approaches when handling the large scale data set on cloud. The centralized TDS approach in [7, 8, 12] perform the data structure TIPS to improve the scalability and efficiency by indexing anonymous data records and retaining statistical information TIPS. The data structure speed up the specialization process because indexing structure avoids frequently scanning entire data sets and storing the statistical result. On the other hand amount of metadata collected to maintain the statistical information and linkage information of record partitions is which compared with data sets. Therefore,

Centralized approaches probably suffer from low efficiency

And scalability when handling large scale data sets. Cloud environment, computation is provisions in the form of Virtual machine (VM).The centralized approach are difficult in handling large scale data sets so on cloud using one single VM if the VM has the highest computation and storage capacity.A distributed TDS approach is proposed to address the anonymization problem which concerns, rather than scalability issues, privacy is one of the main issues so privacy protection against third partiesis also to maintain using anonymization algorithm

A.Top down specialization

The generalization of data is using detailed information in a top-down manner until a K-anonymity is violated. Anonymity is the privacy goal on a combination of attributes. Generally TDS is an iterative process starting from topmost domain values in the taxonomy trees of attributes.

[1]Map Reduce Top down Specialization (MRTDS) generalizes the table by specializingit repetitive starting from the most general state. At each step, a general (i.e. parent) value is specializedinto a specific (i.e. child) value like

Spec:p-->child (p), this process is repeated until specialization start a violation of the anonymity requirement. Thecentralized top down specialization approaches do the indexed partition data structure to improve thescalability and efficiency by indexing anonymous data records and retaining statistical information. But in this approach there is an assumption that all data proposed should fit in memory for the centralized approaches. The amountof metadata retained to maintain the statistical and linkage information of record partitions is compared with larger data set.

B.Two phase top down specilization

Two phase top down approach is perform in TDS in a highly scalable and efficient manner [11] The two phases are based on the two levels of parallelization provisioned by MapReduce on cloud.Job level and task level are the two levels of parallelization of MapReduce on cloud. Job level parallelization means multiple Mapreduce jobs can be executed parallelly to make full use of cloud infrastructure resources. MapReduce becomes more powerful and elastic as cloud can offer infrastructure resources on demand, for example, Amazon Elastic MapReduce service [15]. Task level parallelization refers to multiple mapper/reducer tasks in a MapReduce job are executed simultaneously over data splits. It achieves high scalability by parallelizing multiple jobs on data partitions in the first phase, but the resultant anonymization levels are not identical. To obtain finally consistent anonymous data sets, the second phase is necessary to integrate the intermediate results and anonymize entire data sets. In the first phase, an original data set D is partitioned into smaller ones. Then each of the partitioned data sets in parallel to make full use of the job level parallelization of MapReduce. The subroutine is a MapReduce version of centralized TDS (MRTDS) which specially conducts the computation required in TPTDS. Two Phase MapReduce Top Down Specialization (TPMRTDS) anonymizes data partitions to generate intermediate anonymization levels. An intermediate anonymization level means that further specialization can be performed without violating k-anonymity.MRTDS only advantage the task level parallelization of MapReduce. In the second phase, all intermediate anonymization levels are merged into one. The basic idea of TPTDS is to gain high scalability by making a tradeoff between scalability and data utility. The slight decrease of data utility can lead to high scalability.

3.PROPOSED SYSTEM

In proposed system, a Two-Phase Top-Down Specialization (TPTDS) approach to conduct the computation required in TDS in a highly scalable and efficient manner. The two phases of our approach are based on the two levels of parallelization provisioned by Map Reduce on cloud. Basically, Map Reduce on cloud has two levels of parallelization, i.e., job level and task level.Job level parallelization means that multiple Map Reduce jobs can be executed simultaneously to make full use of cloud infrastructure resources Combined with cloud.Using data partitioning we can partition the data and apply anonymization technique and maintain the privacy after merging technique, Map Reduce becomes more powerful and elastic as cloud can offer infrastructure resources on demand.and using anonymization technique we can keep the identity privacy.The proposed system anonymizes the data by partitioning the attributes and applying proper map-reduce framework on the Hadoop Distributed file system. These approaches get input data's and split into the small data sets. Then we applied the anonymization on small data sets to get intermediate result. Then small data sets are merge and again applied the anonymization. Here the draw back of proposed system is there is no priority for applying the anonymization on datasets. So that its take more time to anonymize the datasets. So we introduce the scheduling mechanism called OPTIMIZED BALANCED SCHEDULING (OBS) to apply the Anonymization. Here the OBS means individual dataset have the separate sensitive field. We analyzed the each and every data set sensitive field and give priority for this sensitive field. Then we apply anonymization on this sensitive field only depending upon the scheduling.

A.Data Partition

When D is partitioned into Di, 1 ≤ i ≤ p,it is required that the data distribution of data records in Di is similar to D. A data record here can be treated as a point in an *m*-dimension space, where *m* is the number of attributes.Thus, the intermediate anonymization levels derived from 1 ≤ i ≤ p, can be more similar so get a better combined Anonymization level. Arbitrary sampling procedure is adopted to partition D, which can satisfy the above condition. Specifically, an arbitrary number rand1≤ rand ≤ p is created for each data record. A record is dispersed to the partition D rand. Data partition map and reduce algorithm the data partition the number of Reducers should be equivalent to P, so that each Reducer handles one value of rand, exactly producing p resultant files. Each file contains a random sample of D.

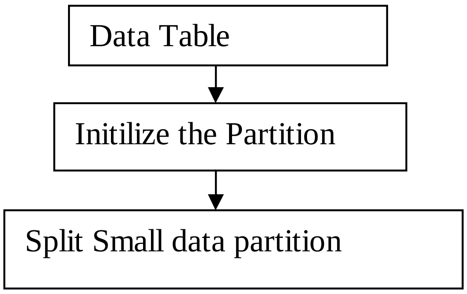
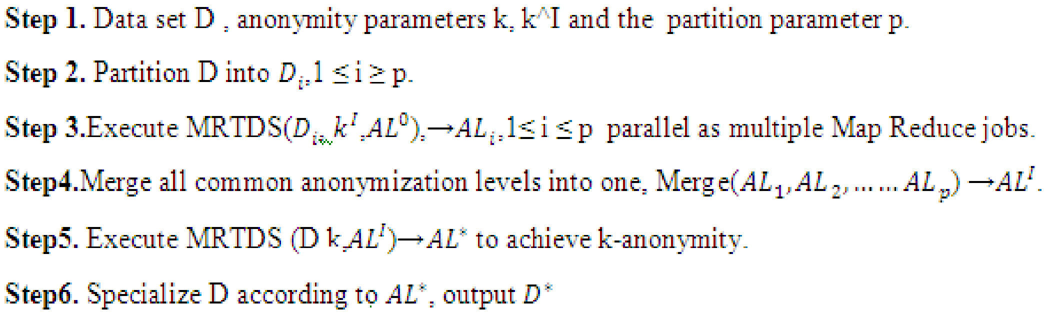


Fig1. Data Partition

Processing data on the storage helps to decrease transmission of data. In Mapping the master node takes the input and divides it into smaller sub group and distributes them into worker nodes.Worker node processes the smaller group and result send back to the master node.In reduce step the master node collects all the results from sub group and integrate them in some way to form the output.So below some steps for data partition using Map and reduce.

Algorithm :- Data Partition Map & Reduce



Step 1. Data set D , anonymity parameters k, k^I and the partition parameter p.

Step 2. Partition D into D_i,1 ≤ i ≥ p.

Step 3. Execute MRTDS(D_i,k^I,AL^0),→AL_i,1≤ i ≤ p parallel as multiple Map Reduce jobs.

Step4. Merge all common anonymization levels into one, Merge(AL_1,AL_2,……AL_p) →AL^I.

Step5. Execute MRTDS (D k,AL^I)→AL^* to achieve k-anonymity.

Step6. Specialize D according to AL^*, output D^*

1. **Prepare the Map() input** – the "MapReduce system" designates Map processors, assigns the K1 input key value each processor would work on, and provides that processor with all the input data associated with that key value.

2. **Run the user-provided Map() code** – Map() is run exactly once for each K1 key value, generating output organized sby key values K2.

3. **"Shuffle" the Map output to the Reduce processors** – the MapReduce system designates Reduce processors, assigns the K2 key value each processor would work on, and provides that processor with all the Map-generated data associated with that key value.

4. **Run the user-provided Reduce() code** – Reduce() is run exactly once for each K2 key value produced by the Map step.

5. **Produce the final output** – the MapReduce system collects all the Reduce output, and sorts it by K2 to produce the final outcome.

B. Anonymization Creation

After gets the individual data sets we applies the anonymization.The anonymization means hide or remove the sensitive field in the data sets.Then we get the intermediate result for small data sets. Anonymization the data is important because in some cases, there is no way to stop this information from being available and some data has high value and analyzing it could lead to progress in fields that are highly important, like hospital and education related data.

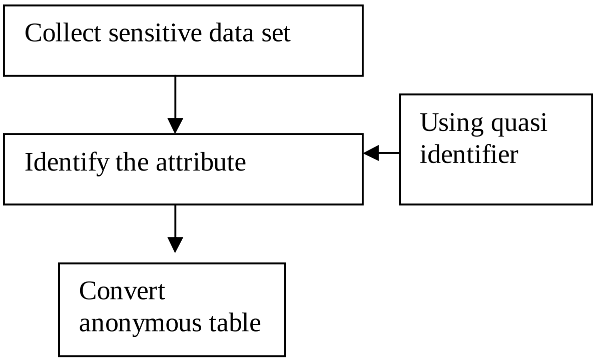


Fig2. Anonymization creation

C.Anonymization level merging

In the second phase of TDS all intermediate anonymization levels merged into one. The merging of anonymization level is completed by merging cuts.This anonymization level of merging denoted as AL.Formally, AL={cut1,cut2,……..cutm}where cuti, 1≤ i ≤ m, is the cut of taxonomy tree TT.

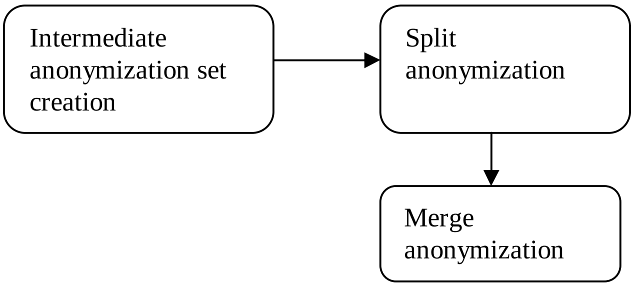


Fig3.Anonymization Merging

D.MRTDS Driver

Usually,a single MapReduce job is deficient to achieve a complex task in many application.MRTDS consist of MRTDS driver and jobs like IGPL Initialization and IGPL Update.

Algorithm : MRTDS Driver

Input : data set D ,anonymization level AL and k-anonymous parameter k.

Step1. Initialize the values of search metric IGPL,i.e for each specialization spec € $\cup_{j=1}^{m} cut_j$. IGPL value of spec is computed by job IGPL initialization.

Step2.While ∃ spec € $\in \cup_{j=1}^{m} cut_j$.. Is valid.

3.1 Find the best specialization from $AL_i, spec_{best}$.

3.2. Update AL_i to AL_{i+1}

3.3.Update information gain of the new specialization in AL_{i+1} , privacy loss for each specialization via job IGPL update.

E.OBS :

The obs called optimized balancing scheduling. Here we focus on the two kinds of the scheduling called time and size. Here data sets are split in to the specified size and applied anonymization on specified time. The obs approach to we provide the high ability on handles the large data sets

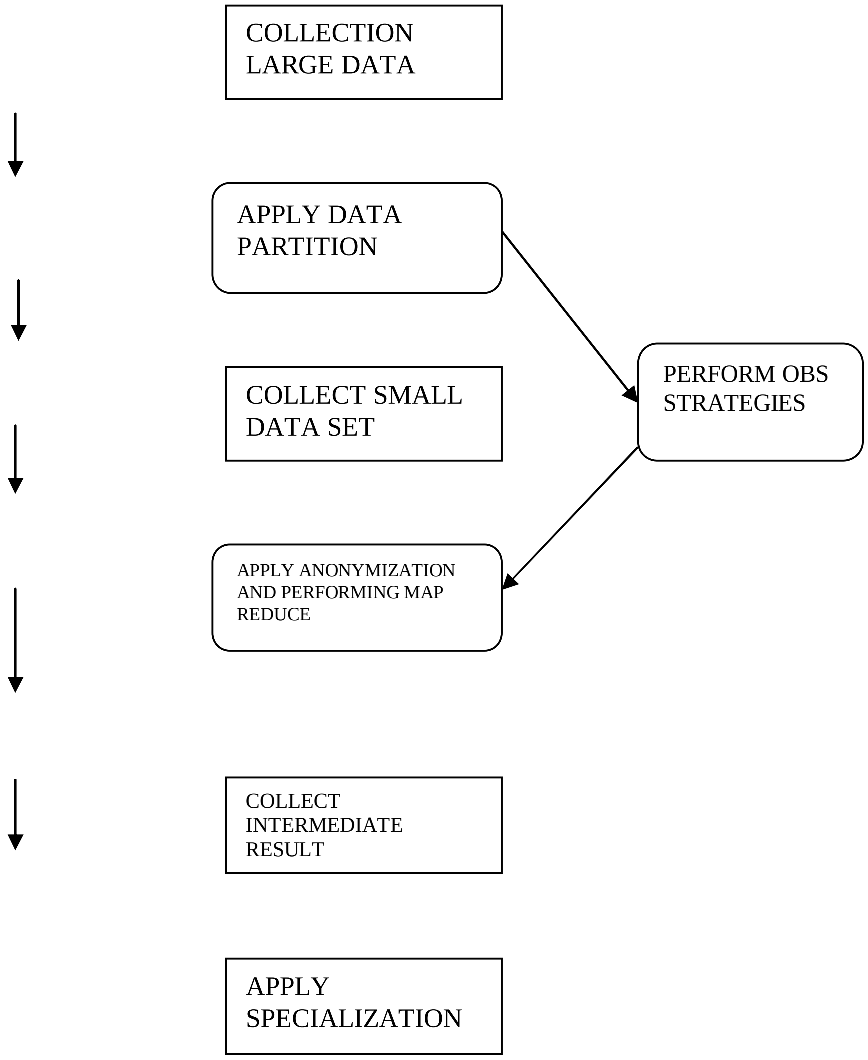


Fig. Data flow diagram of OBS

4.Acknowledgment

I remain immensely obliged to Prof. Harish Baraptre, for providing me with the idea of this topic, and for his invaluable support in garnering resources for me either by way of information or computers also his guidance and supervision which made this happen.

I would like to thank my college Yadavrao Tasgaonkar Institute of Engineering and Technology,Bhivpuri,Karjat and their invaluable support of all staff members.

5.Conclusion

In this paper we have explored the scalability problem of large scale data anonymization by TDS and proposed highly scalable two-phase top down specialization using Map reduce on cloud.In this we proposed anonymized method to preserve the privacy on authorized data.Amenability in publicaly accessible software enables an attacker to break the cloud and display data of other customers using the same service.So considering this issue when we are publicaly store the data on cloud we apply the privacy on sensitive field using anonymization method In this the first phase of TPTDS dataset are partitioned into number of files and anonymized in parallel and producing transitional results.Then in the second phase these results are merged and anonymized to produce consistent k-anonymous data set.

References

[1]. S. Chaudhuri, “What Next?: A Half-Dozen Data Management Research Goals for Big Data and the Cloud,” in 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,” Commun. 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 Distrib. Syst., vol. 23, no. 2, pp.296-303, 2012.

[4]. H. Takabi, J.B.D. Joshi and G. Ahn, “Security and Privacy Callenges in Cloud Computing Environments,”IEEE Security and Privacy, vol. 8, no. 6, pp. 24-31, 2010.

[5]. D. Zissis and D. Lekkas, “Addressing Cloud Computing Security Issues,” Fut. Gener. Comput. Syst., vol. 28, no. 3, pp. 583-592, 2011.

[6]. X. Zhang, Chang Liu, S. Nepal, S. Pandey and J. Chen, “A Privacy Leakage Upper-Bound Constraint Based Approach for Cost-Effective Privacy Preserving of Intermediate Datasets in Cloud,” IEEE Trans. Parallel Distrib. Syst., In Press, 2012.

[7]. 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.

[8]. N. Mohammed, B. Fung, P.C.K. Hung, and C.K. Lee, “Centralized and Distributed Anonymization for High-Dimensional Healthcare Data,” ACM Trans. Knowledge Discovery from Data, vol. 4, no. 4,Article 18, 2010.

[9]. L. Sweeney, “k-Anonymity: A Model for Protecting Privacy,” Int’lJ. Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.

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

[11].Xuyun Zhang, Laurence T. Yang, Senior Member, IEEE, Chang Liu, and Jinjun Chen, “A Scalable Two-Phase Top-Down SpecializationApproach for Data Anonymization Using Map-Reduce on Cloud ”,Feb 2014.

[12].B. Fung, K. Wang, L. Wang, and P.C.K. Hung, “Privacy-Preserving Data Publishing for Cluster Analysis,” Data andKnowledge Eng., vol. 68, no. 6, pp. 552-575, 2009.

[13].N. Mohammed, B.C. Fung, and M. Debbabi, “Anonymity MeetsGame Theory: Secure Data Integration with Malicious Participants,”VLDB J., vol.20, no.4, pp.567-588, 2011.

[14].K. LeFevre, D.J. DeWitt and R.Ramakrishnan, “MondrianMultidimensional K-Anonymity,” Proc. 22nd Int'l Conf. DataEngineering (ICDE '06), article 25, 2006.

[15].D. Zissis and D. Lekkas, “Addressing Cloud Computing Security Issues,” Fut. Gener. Comput. Syst., vol. 28, no. 3, pp.583-592, 2011.