Data Anonymization Approach For Privacy Preserving In Cloud

Saranya M
Computer Science And Engineering
MIET Engineering College
Trichy,India

Senthamil Selvi R
Computer Science And Engineering
MIET Engineering College
Trichy,India

Abstract— Private data such as electronic health records and banking transactions must be shared within the cloud environment to analysis or mine data for research purposes. Data privacy is one of the most concerned issues in big data applications, because processing large-scale sensitive data sets often requires computation power provided by public cloud services. A technique called Data Anonymization, the privacy of an individual can be preserved while aggregate information is shared for mining purposes. Data Anonymization is a concept of hiding sensitive data items of the data owner. A bottom-up generalization approach for transforming more specific data to less specific but semantically consistent data in order to preserve privacy. The idea is to explore the data generalization from data mining to hide detailed data, rather than discovering the patterns. When the data is masked, data mining techniques can be applied without modification.

Keywords—Data Anonymization; Cloud; Bottom Up Generalization; Mapreduce; Privacy Preservation.

I. INTRODUCTION

Cloud Computing refers to configuring, manipulating, and accessing the applications through online. It provides online data storage, infrastructure and application.which is a disruptive trend which poses a significant impact on current IT industry and research communities [1]. Cloud computing provides massive storage capacity computation power and by utilizing a large number of commodity computers together. It enable users to deploy applications with low cost, without high investment in infrastructure. Due to privacy and security problem, numerous potential customers are still hesitant to take advantage of cloud [7].However, Cloud computing reduce costs through optimization and increased operating and economic efficiencies and enhance collaboration, agility, and scale,by enabling a global computing model over the Internet infrastructure. However, without proper security and privacy solutions for clouds, this potentially cloud computing paradigm could become a huge failure.

Cloud delivery models are classified into three. They are software as a service (saas), platform as a service (paas) and infrastructure as a service (iaas). Saas is very similar to the old thin-client model of software provision, clients where usually web browsers, provides the point of access to running software on servers.Paas provides a platform on which software can be developed and deployed. Iaas is comprised of highly automated and scalable computer resources, complemented by cloud storage and network capability which can be metered ,self-provisioned and available on-demand[7].

Cloud is deployed using some models which include public, private and hybrid clouds. A public cloud is one in which the services and infrastructure are provided off-site over the Internet. A private cloud is one in which the services and infrastructure are maintained on a private network. Those clouds offer a great level of security. A hybrid cloud includes a variety of public and private options with multiple providers.

Big data environments require clusters of servers to support the tools that process the large volumes of data, with high velocity and with varied formats of big data. Clouds are deployed on pools of server, networking resources , storage and can scale up or down as needed for convenience.Cloud computing provides a cost-effective way for supporting big data techniques and advanced applications that drives business value. Big data analytics is a set of advanced technologies designed to work with large volumes of data. It uses different quantitative methods like computational mathematics, machine learning, robotics, neural networks and artificial intelligence to explore the data in cloud.

In cloud infrastructure to analyze big data makes sense because Investments in big data analysis can be significant and drive a need for efficient and cost-effective infrastructure, Big data combines internal and external sources as well as Data services that are needed to extract value from big data[17].

To address the scalability problem for large scale data set used a widely adopted parallel data processing framework like Map Reduce. In first phase, the original datasets are partitioned into group of smaller datasets.Now those datasets are anonymized in parallel producing intermediate results. In second phase, the obtained intermediate results are integrated into one and further anonymized to achieve consistent k-anonymous dataset.

Mapreduce is a model for programming and Implementing for processing and generating large data items. A map function that processes a key-value pair,This generates a set of intermediate key-value pair. A reduce function which merges all intermediate data values associated with those intermediate key.

II. RELATED WORK

Ke Wang, Philip S. Yu , Sourav Chakraborty adapts an bottom-up generalization approach which works iteratively to generalize the data. These generalized data is useful for classification.But it is difficult to link to other sources. A hierarchical structure of generalizations specifies the generalization space.Identifying the best generalization is the key to climb up the hierarchy at each iteration[2].

Benjamin c. M. Fung, ke wang discuss that privacy-preserving technology is used to solve some problems only,But it is important to identify the nontechnical difficulties and overcome faced by decision makers when deploying a privacy-preserving technology. Their concerns include the degradation of data quality, increased costs , increased complexity and loss of valuable information. They think that cross-disciplinary research is the key to remove these problems and urge scientists in the privacy protection field to conduct cross-disciplinary research with social scientists in sociology, psychology, and public policy studies[3].

Jiuyong Li,Jixue Liu , Muzammil Baig , Raymond Chi-Wing Wong proposed two classification-aware data anonymization methods .It combines local value suppression and global attribute generalization. The attribute generalization is found by the data distribution, instead of privacy requirement. Generalization levels are optimized by normalizing mutual information for preserving classification capability[17].

Xiaokui Xiao Yufei Tao present a technique,called *anatomy*, for publishing sensitive datasets. Anatomy is the process of releasing all the quasi-identifier and sensitive data items directly in two separate tables. This approach protect the privacy and capture large amount of correlation in microdata by Combining with a grouping mechanism. A linear-time algorithm for computing anatomized tables that obey the l-diversity privacy requirement is developed which minimizes the error of reconstructing microdata [13].

III. PROBLEM ANALYSIS

The centralized Top Down Specialization (TDS) approaches exploits the data structure to improve scalability and efficiency by indexing anonymous data records. But overheads may be incurred by maintaining linkage structure and updating the statistic information when date sets become large.So,centralized approaches probably suffer from problem of low efficiency and scalability while handling large-scale data sets. A distributed TDS approach is proposed to address the anonymization problem in distributed system.It concentrates on privacy protection rather than scalability issues.This approach employs information gain only, but not its privacy loss. [1]

Indexing data structures speeds up the process of anonymization of data and generalizing it, because indexing data structure avoids frequently scanning the whole data[15]. These approaches fails to work in parallel or distributed environments such as cloud systems since the indexing structures are centralized. Centralized approaches are difficult in handling large-scale data sets well on cloud using just one single VM even if the VM has the highest computation and storage capability.

Fung et.al proposed TDS approach which produces an anonymize data set with exploration problem on data. A data structure taxonomy indexed partition [TIPS] is exploited which improves efficiency of TDS, it fails to handle large data set. But this approach is centralized leasing to in adequacy of large data set.

Raj H, Nathuji R, Singh A, England P proposes cache hierarchy aware core assignment and page coloring based cache partitioning to provide resource isolation and better resource management by which it guarantees security of data during processing.But Page coloring approach enforces the performance degradation in case VM’s working set doesn’t fit in cache partition[14].

Ke Wang , Philip S. Yu considers the following problem.Data holder needs to release a version of data that are used for building classification models.But the problem is privacy protection and wants to protect against an external source for sensitive information.So by adapting the iterative bottom-up generalization approach to generalize the data from data mining.

IV. METHODOLOGY

Suppression: In this method, certain values of the attributes are replaced by an asterisk '*'. All or some values of a column may be replaced by '*'

Generalization: In this method, individual values of attributes are replaced by with a broader category. For example, the value '19' of the attribute 'Age' may be replaced by ' ≤ 20', the value '23' by '20 < Age ≤ 30'.

A. Bottom-Up Generalization

Bottom-Up Generalization is one of the efficient k-anonymization methods. K-Anonymity where the attributes are suppressed or generalized until each row is identical with at least k-1 other rows. Now database is said to be k-anonymous. Bottom-Up Generalization (BUG) approach of anonymization is the process of starting from the lowest anonymization level which is iteratively performed. We leverage privacy trade-off as the search metric. Bottom-Up Generalization and MR Bottom up Generalization (MRBUG) Driver are used.The following steps of the Advanced BUG are ,they are data partition, run MRBUG Driver on data set, combines all anonymization levels of the partitioned data items and then apply generalization to original data set without violating the k-anonymity.

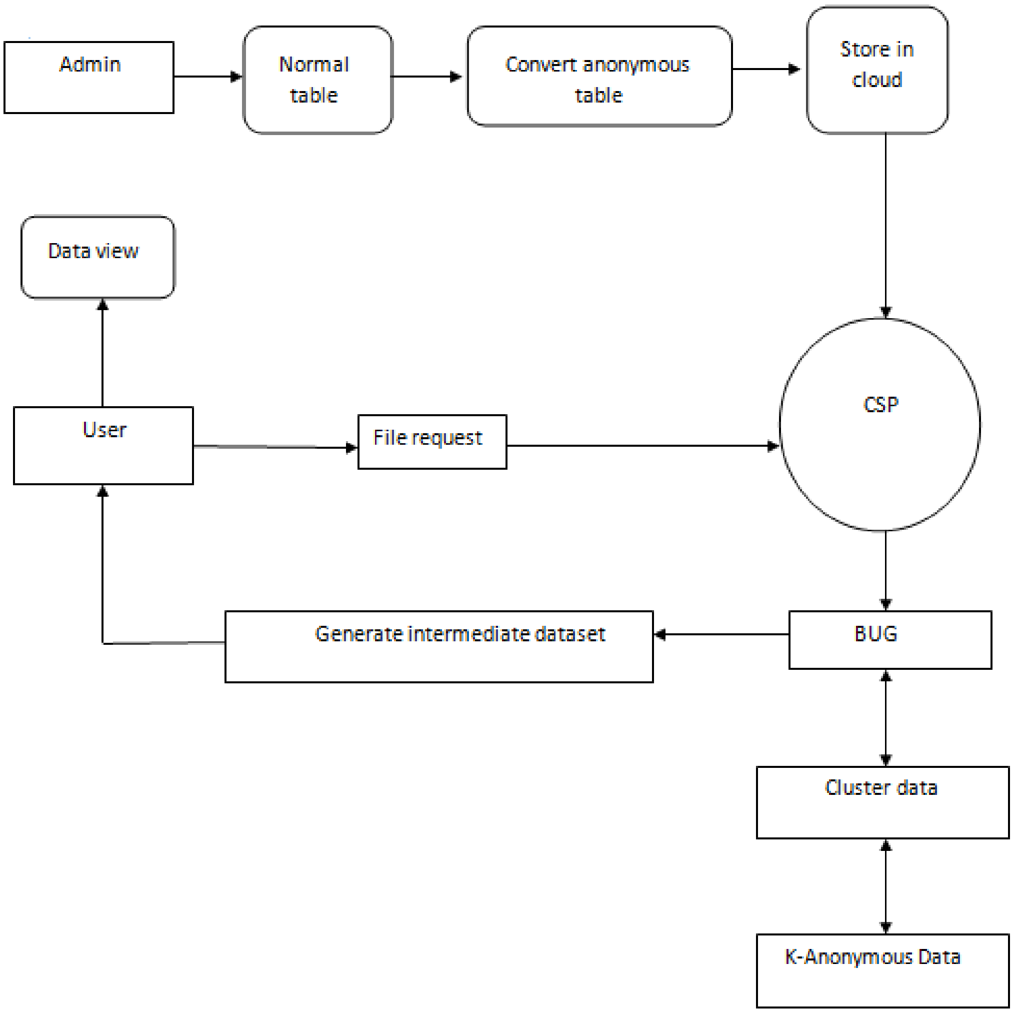

Figure 1. System architecture of bottom up approach

Here an Advanced Bottom-Up Generalization approach which improves the scalability and performance of BUG. Two levels of parallelization which is done by mapreduce(MR) on cloud environment. Mapreduce on cloud has two levels of parallelization.First is job level parallelization which means multiple MR jobs can be executed simultaneously that makes full use of cloud infrastructure.Second one is task level parallelization which means that multiple mapper or reducer tasks in a MR job are executed simultaneously on data partitions.The following steps are performed in our approach,

```
while R that does not satisfy anonymity requirement do
    for all generalizations G do
        compute the IP(G);
    end for;
    find best generalization Gbest;
    generalize R through Gbest;
end while;
output R;
```

Figure 2. Bottom Up generalization algorithm

First the datasets are split up into smaller datasets by using several job level mapreduce, and then the partitioned data sets are anonymized Bottom up Generalization Driver. Then the obtained intermediate anonymization levels are Integrated into one. Ensure that all integrated intermediate level never violates K-anonmity property. Obtaining then the merged intermediate anonymized dataset Then the driver is executed on original data set, and produce the resultant anonymization level.The Algorithm for Advanced Bottom Up Generalization[15] is given below ,

The above algorithm describes bottom-up generalization.In *i*th iteration, generalize *R* by the best generalization *Gbest* .

B. Mapreduce

The Map framework which is classified into map and reduce functions.Map is a function which parcels out task to other different nodes in distributed cluster.Reduce is a function that collates the task and resolves results into single value.

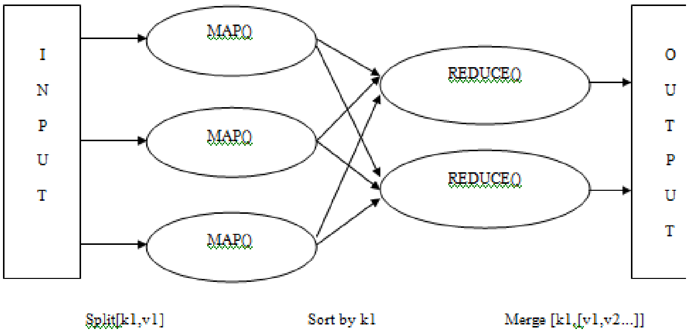

Figure 3. MapReduce Framework

The MR framework is fault-tolerant since each node in cluster had to report back with status updates and completed work periodically.For example if a node remains static for longer interval than the expected,then a master node notes it and re-assigns that task to other nodes.A single MR job is inadequate to accomplish task. So, a group of MR jobs are orchestrated in one MR driver to achieve the task. MR framework consists of MR Driver and two types of jobs.One is IGPL Initialization and other is IGPL Update. The MR driver arranges the execution of jobs.

Hadoop which provides the mechanism to set global variables for the Mappers and the Reducers. The best Specialization which is passed into Map function of IGPL Update job.In Bottom-Up Approach, the data is initialized first to its current state.Then the generalizations process are carried out *k* -anonymity is not violated. That is, to climb the Taxonomy Tree of the attribute till required Anonymity is achieved.

V. Experiment Evaluation

To explore the data generalization from data mining in order to hide the detailed information, rather to discover the patterns and trends. Once the data has been masked, all the standard data mining techniques can be applied without modifying it. Here data mining technique not only discover useful patterns, but also masks the private information. To have investigated the scalability problem of large-scale data anonymization by TDS, and proposed a highly scalable two-phase TDS approach using Map-Reduce on cloud. Data sets are partitioned and anonymized in parallel in the first phase, producing intermediate results. Then, the intermediate results are merged and further anonymized to produce consistent k-anonymous data sets in the second phase. Creatively applied MapReduce on cloud to data anonymization and deliberately designed a group of innovative MapReduce jobs to concretely accomplish the specialization computation in a highly scalable way. Experimental results on real-world data sets have demonstrated that with our approach, the scalability and efficiency of BUG is improved significantly over existing approaches.

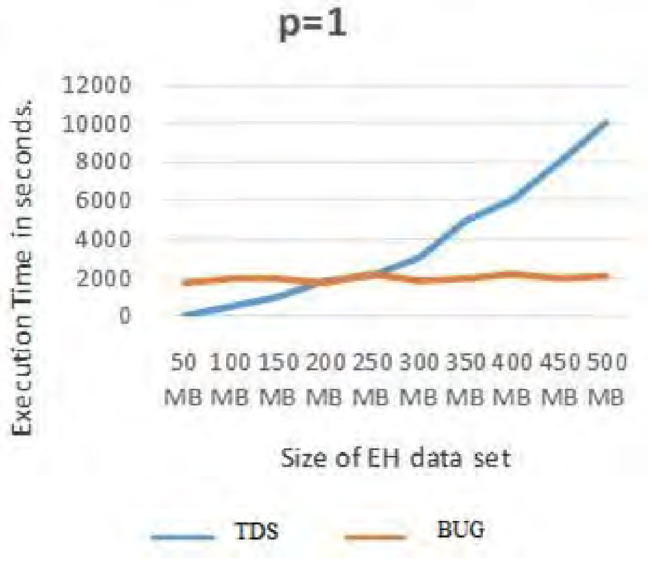

Figure 4.Change of execution time of TDS and BUG

Thus focusing on key issues to achieve quality and scalability. The quality is addressed by trade-off information and privacy and an bottom-up generalization approach.The scalability is addressed by a novel data structure to focus generalizations.To evaluate efficiency and effectiveness of BUG approach, thus we compare BUG with TDS.Experiments are performed in cloud environment.These approaches are implemented in Java language and standard Hadoop MapReduce API.

VI. CONCLUSION

Here the scalability problem for anonymizing the data on cloud for big data applications by using Bottom Up Generalization and proposes a scalable Bottom Up Generalization. The BUG approach performed as follows,first Data partitioning ,executing of driver that produce a intermediate result. After that, these results are merged into one and apply a generalization approach.This produces the anonymized data. The data anonymization is done using MR Framework on cloud.This shows that scalability and efficiency are improved significantly over existing approaches.

REFERENCES

[1] 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”, vol. 25, no. 2, february 2014.

[2] Ke Wang, Yu, P.S,Chakraborty, S, “ Bottom-up generalization: a data mining solution to privacy protection”

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

[4] K. LeFevre, D.J. DeWitt and R. Ramakrishnan, “Workload- Aware Anonymization Techniques for Large-Scale Datasets,” ACM Trans. Database Syst., vol. 33, no. 3, pp. 1-47, 2008.

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

[6] 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.

[7] Hassan Takabi, James B.D. Joshi and Gail-Joon Ahn, “Security and Privacy Challenges in Cloud Computing Environments”.

[8] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, “Incognito: Efficient Full-Domain K-Anonymity,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’05), pp. 49-60, 2005.

[9] T. IwuchukwuandJ.F. Naughton, “K-Anonymization as Spatial Indexing: Toward Scalable and Incremental Anonymization,” Proc. 33rdInt'lConf. VeryLarge DataBases (VLDB'07), pp.746-757, 2007

[10] J. Dean and S. Ghemawat, “Mapreduce: Simplified Data Processing on Large Clusters,” Comm. ACM, vol. 51, no. 1, pp. 107-113,2008.

[11] Dean J, Ghemawat S. “Mapreduce: a flexible data processing tool,” Communications of the ACM 2010;53(1):72–77. DOI: 10.1145/1629175.1629198.

[12] Jiuyong Li, Jixue Liu , Muzammil Baig , Raymond Chi-Wing Wong, “Information based data anonymization for classification utility”

[13] X. Xiao and Y. Tao, “Anatomy: Simple and Effective Privacy Preservation,” Proc. 32nd Int’l Conf. Very Large Data Bases (VLDB’06), pp. 139-150, 2006.

[14] Raj H, Nathuji R, Singh A, England P. “Resource management for isolation enhanced cloud services,” In: Proceedings of the 2009ACM workshop on cloud computing security, Chicago, Illinois, USA, 2009, p.77–84.

[15] K.R.Pandilakshmi, G.Rashitha Banu. “An Advanced Bottom up Generalization Approach for Big Data on Cloud” , Volume: 03, June 2014, Pages: 1054-1059..

[16] Intel “Big Data in the Cloud: Converging Technologies”.

[17] Jiuyong Li, Jixue Liu Muzammil Baig, Raymond Chi-Wing Wong, “Information based data anonymization for classification utility”.