A Review of Scalable Privacy Preservation in Big Data

[1]Priyanka Gawali
[1,]Dr. D. Y. Patil School of Engineering & Technology, Pune, Maharashtra, India.
[1]priyankagawali@gmail.com

Abstract : BigData Application uses Cloud computing environments provides flexible infrastructure and high storage capacity. For processing huge amount of unstructured data in Bigdata applications MapReduce Technique is used. Increase in data volume leads to flexible and scalable privacy preservation of such dataset over the MapReduce framework is BigData applications. A Review has been taken for the MapReduce Technique based big data privacy preservation in Cloud environment. Existing approaches employ local recording anonymization for privacy preserving where data are processed. This processed data used for analysis, mining. The proposed work focus cloud environments on Local recording anonymization for preserving data privacy over BigData using MapReduce.

Keywords — Privacy, BigData, Cloud, MapReduce, Privacy, Scalable.

I. INTRODUCTION

CLOUD computing and BigData, two disruptive trends at present, pose a significant impact on current industry and research community. Today, a large number of big data services are deployed or migrated to cloud for data mining, processing or sharing. The salient characteristics of cloud computing such as high scalability and pay-as-you-go fashion make Big Data inevitably accessible by various organizations through public cloud infrastructure. Data sets in Big Data applications often contain personal privacy sensitive data like electronic health records and financial transaction records. As the analysis of these data sets provides profound insights into a number of key areas of society, the data sets are often shared or released to third party partners or the public. So it is essential for strong preservation of data privacy. Data anonymization plays major role in privacy preservation in non-interactive data sharing and releasing process. Data anonymization[10, 12] refers to hiding identity of sensitive data so privacy of an individual is preserved even certain aggregate information can be still exposed to data users for diverse analysis and mining tasks. A various of privacy models and data anonymization approaches have been proposed and extensively reviewed [5, 8]. However, applying these traditional approaches to big data anonymization poses scalability and efficiency challenges because of the 3Vs, Volume, Velocity and Variety. The research on scalability issues of big data anonymization came to the picture but they lack in some common issues.

II. RELATED WORK

Xuyun Zhang et. al.,[1] have investigated local-recoding anonymization for big data in cloud from the perspective of capability of defending proximity privacy breaches, scalability and time efficiency. A proximity privacy model was proposed against privacy breaches. A scalable two-phase clustering approach based on MapReduce was proposed to address the above problem in time efficiently. Extensive experiments on real-world data sets demonstrates that this paper research approach significantly improves the capability of defending proximity attacks, the scalability and the time efficiency of local-recoding anonymization . Local recording scheme partitions the data set in clustering fashion ,where top-down anonymization

is inapplicable leads to inefficient privacy. this approach tailored for small scale data sets often fall short when encountering BigData. Wanchun Dou et. al.,[2] have enhanced History record-based Service optimization method, named Hire Some-II ,a cross-cloud service composition for processing big data applications. It can effectively promote cross-cloud service composition in the situation where a cloud refuses to disclose all details of its service transaction records for business privacy issues in cross-cloud scenario. This method significantly reduces the time complexity as only some representative history records are recruited, which is highly demanded for BigData applications. of its transaction records, which accordingly protects privacy in big data. Here, the credibility of cross-clouds and on-line service compositions will become suspicioned, if a cloud fails to deliver its services according to its 'promised' quality. Xueli Huang et. al.,[3] proposed an efficient scheme to address the increasing concern of data privacy in cloud for image data. The proposed scheme divides an image into blocks and shuffles the blocks with random start position and random stride which operates at the block level instead of the pixel level, which greatly speeds up the computation The proposed scheme was implemented real networks (including the Amazon EC2)and tested the security and efficiency. Both analysis and experimental results showed that the proposed scheme is secure, efficient but has very small overhead and its only applicable for image data. Unstructured data are out of focus. Jeff Sedayaoet. al.,[4] suggested to use Hadoop to analyse the anonymized data and obtain useful results for the Human Factors analysts. At the same time, the requirements of anonymization were learned and anonymized data sets need to be carefully analysed to determine whether they are vulnerable to attack. Anonymization tools were found intended for the enterprise generally did not seem to consider the quality of anonymization and does not clearly state whether an anonymized data set was vulnerable to correlation attacks. Wenyi Liu et. al.,[5]were developed a privacy-preserving multi factor authentication system without introduction of any extra physical device for cloud systems utilizing big data features has two advantages over previously proposed systems. First, user privacy is not leaked to ubiquitous cloud computing environment .Second , the hybrid user profiling model is highly usable and configurable and integrates a lot of features and corresponding data, which enables simple privacy-preserving operations with fuzzy-hashing calculations. One can always modify the feature list for user profiling according to the actual circumstances. The system performance was evaluated via a series of experiments utilizing four different datasets, resulting in an optimal recall of 80.8%. Also, both system overhead and resource utilization were within the acceptable range, which substantiates the feasibility of the scheme. Adding more features and including a weighting scheme on features that can be configured by the system administrator and plan to improve performance to be considered. Xuyun Zhang et. al.,[7] investigated the scalability issue of multidimensional anonymization over big data on cloud, and proposed a scalable MapReduce based approach. The scalability issues of finding the median due to its core role in multidimensional partitioning was examined and highly scalable MapReduce based algorithm was proposed for finding the median and histogram technique. More number of experiments on datasets were conducted which would be extended from real life datasets, and the experimental results demonstrate that the scalability and cost effectiveness of multidimensional anonymization scheme can be improved significantly over existing approaches. But ensuring privacy preservation of large scale data sets still needs extensive investigation, if this work is integrated into scalable and cost effective privacy preserving framework. Scalable privacy preservation aware analysis and scheduling on big data is to be considered. Meiko Jensen et. al.,[9] ,explained that the field of privacy in big data contexts contains a bunch of key challenges that must be addressed by research. Many of these challenges do not stem from technical issues, but merely are based on legislation and organizational matters. Nevertheless, it can be anticipated that it was feasible to meet each of the challenges discussed here by means of appropriate technical measures.

III. TRADITIONAL DATA PRIVACY PRESERVATION METHODS

Cryptography refers to set of techniques and algorithms for protecting data. In cryptography plaintext is converted into cipher text using various encryption schemes. There are various methods

based on this scheme like public key cryptography, digital signatures etc. Cryptography alone can't enforce the privacy demanded by common cloud computing and big data services [14]. This is because big data differs from traditional large data sets on the basis of three V's (velocity, variety, volume) . It is these features of big data that make big data architecture different from traditional information architectures. These changes in architecture and its complex nature make cryptography and traditional encryption schemes not scalable up to the privacy needs of big data.

The challenge with cryptography is all or nothing retrieval policy of encrypted data. The less sensitive data that can be useful in big data analytics is also encrypted and user is not allowed to access it. It makes data inaccessible to those who don't have access to decryption key. Also privacy may be breached if data is stolen before encryption or cryptographic keys are misused. Attribute based encryption can also be used for big data privacy [14]. This method of securing big data is based on relationships among attributes present in big data. The attributes that need to be protected are identified based on type of big data and company policies. In nutshell, encryption or cryptography alone can't stand as big data privacy preservation method. They can help us to do data anonymization but cannot be used directly for big data privacy.

IV. PRIVACY PRESERVING APPROACHES IN DATA PUBLISHING

4.1. K-ANONYMITY

K-anonymity is a property possessed by certain anonymized data. Given person-specific field-structured data; produce a release of the data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful. A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k-1 individuals.

4.1.1. k-anonymization Methods

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

Generalization: In generalization, individual values of attributes are replaced by with a broader category.

Table 1. *Anonymized Table*

Name	Age	Gender	State of	Religion	Disease
*	20	Female	Maharashtra	*	Cancer
*	20	Female	Gujrat	*	Viral
*	20	Female	Maharashtra	*	TB
*	20	Male	Gujrat	*	No
*	20	Female	Delhi	*	Heart-

4.2.L-DIVERSITY

L-diversity is a form of group based anonymization. L diversity is used to preserve .The l-diversity model is an extension of the k-anonymity model which reduces the granularity of data representation using techniques including generalization and suppression such that any given record maps onto at least k other records in the data.

The l-diversity approach also handles some of the deficiency in the k-anonymity approach where protected identities to the level of k-individuals is not equivalent to protecting the corresponding sensitive values that were generalized or suppressed, especially when the sensitive values within a group exhibit homogeneity.

4.3 T-CLOSENESS

Given the existence of attacks where sensitive attributes may be inferred depend upon the distribution of values for l-diverse data, the t-closeness apprach was created to further l-diversity by additionally maintaining the distribution of sensitive fields.

An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class & distribution of the attribute in the whole table is no more than a threshold t.

A. *Generalization Approach*

By creatively applying Map Reduce on cloud to Bottom Up Generalization (BUG) for data anonymization and deliberately design a group of innovative Map Reduce jobs to concretely accomplish the generalizations in a highly scalable way. Secondly, introduce a scalable Advanced BUG approach, which performs generalization on different partitioned data set and the resulting intermediate anonymizations are

merged to find final anonymization which is used to anoymize the original data set. Results show that our approach can significantly improve the scalability and efficiency of BUG for data anonymization over existing approaches.

4.5 Top-Down Specialization

Generally, TDS is an iterative process starting from the topmost domain values in the taxonomy trees of attributes. Each round of iteration consists of three steps [4] Such a process is repeated until k-anonymity is violated, to expose the maximum data utility. The goodness of a specialization is measured by a search metric.

4.6 MAP REDUCE: A LARGE-SCALE DATA PROCESSING FRAMEWORK

To address the scalability problem of the Top-Down Specialization (TDS) approach for large scale data set used a widely adopted parallel data processing framework like Map Reduce. In first half, the original datasets are partitioned into group of smaller datasets and these datasets are anonymized in parallel producing intermediate results. In second half, these intermediate results are integrated into one and further anonymized to achieve consistent k-anonymous dataset.

Mapreduce is used to split up the large input data into chunks of more or equal size, spinning up a number of processing instances for the map phase apportioning data to each of the mappers, tracking the status of each mapper, routing the map results to the reduce phase and finally shutting down the mappers and the reducers when the work has been done. It is easy to scale up MapReduce Framework to handle bigger jobs or to produce results in a shorter time by simply running the job on a larger cluster. When Mapreduce Framework is not used the process fails in distribution system.

4.7 TWO-PHASE TOP-DOWN SPECIALIZATION (TPTDS)

A TPTDS approach in TDS is a highly scalable and efficient approach. 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

Job level parallelization deals multiple MapReduce jobs that can be executed simultaneously to make full use of cloud infrastructure resources. Combined with cloud, MapReduce becomes more powerful and elastic as cloud can offer infrastructure resources on demand, for example, Amazon Elastic MapReduce service [13].

Task level parallelization refers to that multiple mapper/reducer tasks in a MapReduce job are executed simultaneously over data splits. By parallelizing multiple jobs on data partitions in the first phase to achieve high scalability,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 further anonymize entire data sets.

V. SYSTEM DESCRIPTION

Currently, more number of security approaches is available in big data for local recording anonymization. Separate methods were used in existing work. Only with a limited number of verification big data approaches are available. There is no System verification for Big data using MapReduce, Data processing and privacy preserving for global recording anonymization. The proposed work schemes new algorithm for MapReduce in big data for global recording anonymization. If, integration of MapReduce, a tool for privacy preserving, for the analyzing of data is used, it will provide better privacy in scalable big data during uncertain condition. In this section, introduces a two phase top-down specialization approach and it introduce the scheduling mechanism called Optimized Balanced Scheduling(OBS) to apply the anonymization. Here the OBS means individual dataset have the separate sensitive field.

A. Privacy Preserving Map-Reduce Cloud

The main drawback of the approach is centralized top-down approach. It's does not have the ability for handle the large scale datasets in cloud. Its overcome by it invent the two phase top-down specialization approach. This approach gets input data's and split into the small data sets. Small data

sets are merging, then its uses for 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 it introduces the scheduling mechanism called Optimized Balanced Scheduling (OBS) to apply the anonymization. Here the OBS means individual dataset have the separate sensitive field. It analyzes the each and every data set sensitive field and gives priority for this sensitive field. Then apply anonymization on this sensitive field only depending upon the scheduling.

B. Two Phase Top Down Specialization

Two-Phase Top-Down Specialization (TPTDS) approach to conduct the computation required in TDS in a highly scalable and efficient fashion. The two phases of the approach are based on the two levels of parallelization provisioned by MapReduce on cloud. Basically, MapReduce on cloud has two levels of parallelization, i.e., job level and task level. Job level parallelization means that multiple MapReduce jobs can be executed simultaneously to make full use of cloud infrastructure resources. Combined with cloud, MapReduce becomes more powerful and elastic as cloud can offer infrastructure resources on demand, e.g., Amazon Elastic MapReduce service. Task level parallelization refers to that multiple mapper/reducer tasks in a MapReduce job are executed simultaneously over data splits. To achieve high scalability, 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 further anonymize entire data sets. Details are formulated as follows. All intermediate anonymization levels are merged into one in the second phase. The merging of anonymization levels is completed by merging cuts. Specifically, let in and in be two cuts of an attribute. There exist domain values and that satisfy one of the three conditions is identical to is more general than is more specific than. To ensure that the merged intermediate anonymization level never violates privacy requirements, the more general one is selected as the merged one, e.g., will be selected if is more general than or identical to . For the case of multiple anonymization levels, it can merge them in the same way iteratively. The following lemma ensures that still complies privacy requirements.

VI. CONCLUSION

Currently, security in Big data is a challenging research issue. If Integration of MapReduce, a machine for privacy preserving, is designed for the analyzing of data would provide better privacy. In the existing system scalability and time-efficiency have been done with local-recording anonymization and did not address global-recording anonymization. This review work gives idea Local recording anonymization in cloud environments for preserving data privacy over BigData using MapReduce. Using the two phase top down approach to provide ability to handles the high amount of the large data sets. And here it provides the privacy by effective anonymization approaches.

REFERENCES

[1] Xuyun Zhang, Wanchun Dou, Jian Pei, Surya Nepal, Chi Yang, Chang Liu, and Jinjun Chen, " Proximity-Aware Local-Recoding Anonymization with MapReduce for Scalable Big Data Privacy Preservation in Cloud" in press, 200x(In press).

[2] Wanchun Dou, Xuyun Zhang, Jianxun Liu, and Jinjun Chen," HireSome-II: Towards Privacy-Aware Cross-Cloud Service Composition for Big DataApplications", pp.1-14, 2013.

[3] Xueli Huang and Xiaojiang Du, "Achieving Big Data Privacy via Hybrid Cloud", IEEE INFOCOM Workshops: pp.512-517, 2014.

[4] Jeff Sedayao, Rahul Bhardwaj and NakulGorade, "Making Big Data, Privacy, and Anonymization work together in the Enterprise:Experiences and Issues", IEEE International Congress on Big Data, pp.1-7, 2014.

[5] Wenyi Liu, A. SelcukUluagac, and RaheemBeyah, "MACA: A Privacy-Preserving Multi-factor Cloud Authentication System Utilizing Big Data", IEEE INFOCOM Workshops, pp. 518- 523, 2014.

[6] Amine Rahmani, Abdelmalek Amine, Reda Mohamed Hamou, "A Multilayer Evolutionary Homomorphic Encryption Approach for Privacy Preserving over Big Data", Proceedings of International Conference on

Cyber-Enabled Distributed Computing and Knowledge Discovery , pp. 19-26, 2014.

[7] Xuyun Zhang, Chi Yang, Surya Nepal, Chang Liu, Wanchun Dou, Jinjun Chen, "A MapReduce Based Approach of Scalable Multidimensional Anonymization for Big Data Privacy Preservation on Cloud", Proceedings of 3rdInternational Conference on Cloud and Green Computing, IEEE, pp. 105-112, 2013.

[8] Omar Hasan, Benjamin Habegger, Lionel Brunie, Nadia Bennani, Ernesto Damiani, "A Discussion of Privacy Challenges in User Profiling with Big Data Techniques: The EEXCESS Use Case", IEEE International Congress on Big Data , pp. 1-6, 2013.

[9] Meiko Jensen and Kiel, "Challenges of Privacy Protection in Big Data Analytics", Proceedings of International Congress on Big Data, IEEE, pp. 235- 238, 2013.

[10] AntorweepChakravorty, Tomasz Wlodarczyk, Chunming Rong, "Privacy Preserving Data Analytics for Smart Homes", IEEE Security and Privacy Workshops, pp. 1-5, 2013.

[11] Koichiro Hayashi and Yokohama, "Social Issues of Big Data and Cloud:Privacy, Confidentiality, and Public Utility", Proceedings of 8th International Conference on Availability, Reliability and Security, pp. 506-511, 2013.