ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

A Survey on Privacy Conservation over Big Data

Arundhati Reddy[1], Prof. Mininath K. Nighot[2]

Student, Department of CE, K J College of Engg. & Management Research, Savitribai Phule Pune University, India[1]

Professor, Department of CE, K J College of Engg. & Management Research, Savitribai Phule Pune University, India[2]

ABSTRACT: A BigData Application use Cloud computing environments provides flexible infrastructure and high storage capacity. MapReduce Technique is used for processing large amount of unstructured data in Big data applications. Increase in data volume leads to flexible and scalable secrecy conservation of such dataset over the MapReduce framework is BigData applications. A Review has been taken for the MapReduce Technique based big data privacy conservation in Cloud environment. Available existing overtures employ local recording anonymization for privacy conserving where data are processed. Processed data used for analysis, mining. The proposed work focus cloud environments on Local recording anonymization for conserving data privacy over BigData using MapReduce.

KEYWORDS: BigData, Privacy, MapReduce, Scalable, Anonymization .

I. INTRODUCTION

Bigdata and cloud computing, two disruptive trends at present. Both poses a important impact on current industry and research community. Today, a large number of big data services are migrated to cloud for data mining, processing or sharing. The salient high scalability and pay as you go characteristics of cloud computing make Big Data inevitably accessible by multiple organizations through public cloud infrastructure. Data sets in Big Data applications often contain personal private sensitive data like electronic health records and financial transaction records etc. As the analysis of these data sets provides profound perceptiveness into a number of main areas of society, the data sets are often shared or released to third party or the public. So it is essential for strong conservation of data privacy. Data anonymization plays major role in privacy conservation in non-interactive data sharing and exiting process.
Data anonymization refers to hiding identity of sensitive data so privacy of an individual is conserved even certain aggregate information can be still exposed to data users for various analysis and mining tasks. Several of privacy models and data anonymization overtures have been highly reviewed. However, applying these traditional overtures to big data anonymization poses scalability and efficiency difficulties 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 usual issues.

II. RELATED WORK

Xuyun Zhang et. al.,[1] have identified local-recoding anonymization for big data in cloud from the perspective of capability of supporting proximity privacy breaks, scalability and time efficiency. A proximity privacy model was proposed against privacy breaks. A scalable two-phase clumping overture based on MapReduce was proposed to address the problem in time efficiently. Extensive experiments on real-world data sets shows that this paper research overture importantly improves the capability of supporting proximity attacks, the time efficiency and the scalability of local-recoding anonymization . The Local recording scheme separations the data set in clumping fashion, where top-down anonymization is inapplicable leads to inefficient privacy. This overture tailored for small scale data sets often fall short when encounters BigData.
Wanchun Dou et. al.,[2] have enhanced History record-based Service optimisation method, named HireSome-II , cross-cloud service composition, for processing big data applications. It can effectively advertise cross-cloud service composition in the situation where a cloud refuses to open all details of its service transaction records for business privacy issues in cross-cloud scenario. This method mainly reduces the time complexness as only some representative

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

history records are recruited, which is highly needed for BigData applications. Of its transaction records, which accordingly guard privacy in big data. Here, the credibility of cross-clouds and on-line service compositions will become doubtful, if a cloud fails to deliver quality services according to its promise.

Xueli Huang et. al.[3] proposed an efficient scheme to address the increasing fear of data privacy in cloud for image data. The proposed scheme splits an image into blocks and shuffles the blocks with random initial position and random step which operates at the block level alternative of the pixel level, which greatly speeds up the computation. The proposed scheme was enforced real networks (including the Amazon EC2) and tested the privacy and efficiency. Both analysis and experimental results showed that the proposed scheme is assure, efficient but has very small overhead and it's only applicable for image data. Not structured data are not in focus.

Jeff Sedayaoet. al.,[4] suggested to use Hadoop to analyse the anonymized data and obtain useful results for the Human Factors analyst. At the same time, the want of anonymization were learned and anonymized data sets want to be carefully analysed to determine whether they are exposable to attack. Anonymization tools were found mean for the enterprise generally did not appear to consider the quality of anonymization and does not specify clearly state whether an anonymized data set was week to correlation attacks. Wenyi Liu ET. Al ,[5] were developed a privacy-conserving multi factor authentication system without introduction of any more physical device for cloud systems utilizing big data features has two advantages over proposed systems. First, user privacy is not leaked to ubiquitous cloud computing environment .Second , the hybrid user profiling model is more usable and configurable as well as integrates a lot of features and corresponding data, which enables simple privacy-conserving operations with fuzzy-hashing calculations. One can always alter the feature list for user profiling according to the real circumstances. The system performance was evaluated through series of experiments utilizing four different datasets, resulting in an maximal recall of 80.8%. Also, both system overhead and resource utilization were inside the acceptable range, which substantiates the feasibleness of the scheme. Adding more characteristics and including a weighting scheme on features that can be designed by the system administrator and plan to improve performance to be considered. In [6] Amine Rahmani, Abdelmalek Amine, Reda Mohamed hamou explained about one of the biggest thing is the privacy of individual users. On other side a new type of cryptosystems was recently introduced and his aim is to introduce and improve old cryptography techniques. In his paper he improves the efficiency of homomorphic cryptosystem.

Xuyun Zhang et. al.,[7] enquired the scalability issue of multidimensional anonymization over big data on cloud, and proposes a scalable MapReduce based overture. The scalability issues of finding the average due to its main role in multidimensional partitioning were examined and highly scalable. MapReduce based algorithm was proposed for finding the average and histogram technique. More number of experiments on datasets were carried out which would be extended from real life datasets, and the experimental results shows that the scalability and cost effectiveness of multidimensional anonymization scheme can be meliorate importantly over existing overtures. But ensuring privacy conservation of large scale data sets still needs deep investigation, if this work is integrated into scalable and cost effective privacy conserving structure. Meiko Jensen et. al.,[8] ,elaborated that the field of secrecy in big data contexts contains a bunch of important challenges that must be addressed by research. Many of these challenges do not stalk from technical issues, but merely are depend on legislation and organizational matters. Nevertheless, it can be anticipated that it was workable to meet each of the challenges discussed here by means of advantageous technical measures.

In [9] AnjanaGosain, Nikita Chugh they explained about Big data has brought a revolution in this world of data Analytics. Data that was not considered a few years back is now considered a powerful asset. Big data is being extensively used for knowledge discovery in now a day by all sectors of society. It is generated by almost all digital processes, is stored and shared on web. This reliance of big data on web model poses serious privacy concerns. Scalable privacy [10] conservation aware analysis and scheduling on big data is to be considered.

III. TRADITIONAL DATA PRIVACY CONSERVATION METHODS

Cryptography refers to set of techniques and algorithms for securing data. In cryptography plaintext is converted into cipher text using assorted coding schemes. There are various methods based on this scheme like public key cryptography, digital signatures etc. Cryptography only can't enforce the privacy demanded by common cloud computing and big data services.[9] This is because big data differs from traditional large data sets on the basis of three V's (velocity, variety, volume).These features of big data make big data architecture different from old information

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

architectures. These changes in architecture and its complexity make cryptography and traditional coding schemes not scalable up to the privacy needs of big data.
The challenge with cryptography is all or nothing recovery policy of coded data. The less sensitive data that can be useful in big data analytics is also coded and user is not allowed to access. It makes data inaccessible to those who don‘t have access to decryption key. Also privacy may be breached if data is stolen before coding or cryptographic keys are misused. Attribute based coding can also be used for big data privacy [9] . This method of securing big data is based on relationships among ascribes present in big data. This describes that need to be protected are identified founded on type of big data and company policies. In nutshell, coding or cryptography alone can‘t stand as big data privacy conservation method. They can help us to do data anonymization but can be used alternatively for big data privacy.

IV. PRIVACY CONSERVING OVERTURES IN DATA PUBLISHING

A. K-ANONYMITY

K-anonymity is a characteristics possessed by some specific anonymized data. Given person special field structured data; produce are lease of the data with scientific assuerities that the individuals who are the subjects of the data cannot be reidentified while the data remains practically very helpful. A release of data is said to have the k-anonymity property if the information for each person contained in the release can’t be categorised from at least k-1individuals.

a. k-anonymization Methods

Suppression: In suppression, certain values of the inputes are replaced by an asterisk mark *. All or some values of a column may be interchanged by*, refer Table 1. Generalization: In generalization, individual values of inputs are replaced by with a broader category.

Table1. Anonymized Table

Name	Age	Gender	State	Religion	Disease
*	20 <	Female	Gujrat	*	cancer
*	20 <	Female	Maharashtra	*	TB
*	20 <	Male	Gujrat	*	Noillness
*	20 <	Female	Delhi	*	Heart-related

B.L-DIVERSITY

It is a form of group based anonymization. L diversity is used to conserve. The l-diversity model is an expansion of the k-anonymity model which reduces the granularity of data representation using old or new techniques including generalization and suppression such that any given record maps on to at least k other records in the data.
The l-diversity overture also helps in some of the deficiency in the k-anonymity overture where protected identities to the level of k-individuals is not equivalent to protecting the corresponding sensitive data that were generalized or suppressed, especially when the sensitive values within a group.

C. T-CLOSENESS

Given the existence of assaults where sensitive imputes may be inferred depend upon the distribution of values for l-diverse data; the t-closeness overture was generated to further l-diversity by additionally managing the distribution of sensitive fields. An equivalence class is said to have t-closeness if the distance between the distributions of a sensitive impute in this class &distribution of the attribute in the whole table is no more than a threshold t.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

D. GENERALIZATION OVERTURE

By creatively applying MapReduce on cloud to Bottom Up Generalization (BUG) for data anonymization and deliberately structure a group of innovative MapReduce jobs to concretely carry the generalizations in a highly scalable way. Secondly, introduce a scalable Advanced BUG overture, which performs generalization on different partitioned data set and the resulting intermediate anonymization are united to find final anonymization which is used to anoymize the original data set. Results show that our overtures can efficiency of BUG for data anonymization over existing overtures.

E. TOP-DOWN SPECIALIZATION

Generally, TDS is an iteration process starting from the top domain values in the taxonomy trees of imputes. Each round of iteration consists of three steps such a process is repeated until k-anonymity is violated, to expose the maximum data utilisation. The goodness of a specialization is measured by a search metric.

F. MAP REDUCE: A LARGE-SCALE DATA PROCESSING STRUCTURE

To address the scalability problem of the Top-Down Specialization (TDS) overture for large scale data set used a widely uses parallel data processing structure like MapReduce. In first half, the original data sets are saperated into group of smaller datasets and these datasets are anonymized in parallel producing in term results. In second , these mediator results are integrated into one and further anonymized to achieve consistent k-anonymous dataset. Mapreduce is used to partition large input data into chunks of more or equal size, spinning up a number of processing instances for the map phase allotment data to each of the mappers, tracking the status of each mapper, routing the map results to their duce phase and finally closing down the mappers and the reducers when the work has been done. It is easy to scale up MapReduce structure to handle bigger jobs or to generate results in a shorter time by simply running the job on a larger cluster. When Mapreduce structure is not used the process fails in distribution system.

G. TWO-PHASE TOP-DOWN SPECIALIZATION(TPTDS)

A TPTDS overture in TDS is a highly scalable and effective overture. The two phases of our overture are based on the two levels of parallelization provisioned by MapReduce on cloud. Basically, Map Reduce on cloud has two levels of parallelization.
Job level parallelization deals multiple MapReduce jobs that can be carry out simultaneously to make full use of cloud infrastructure resources. Concerted with cloud, MapReduce becomes more powerful and elastic as cloud can offer infrastructure resources on request, for example, Amazon Elastic MapReduce service.
Task level parallelization refers to that multiple mapper/reducer tasks in a MapReduce job are executed at the same time over data splits. By parallelizing multiple jobs on data separations in the first phase to achieve high scalability, but the resultant anonymization levels different. To obtain finally consistent anonymous data sets, the second case is necessary to incorporate the intermediate results and further anonymize entire datasets.

V. SYSTEM DESCRIPTION

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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

A. PRIVACY CONSERVING MAP-REDUCE CLOUD

The main drawback of the overture is centralized top-down overture. It‘s does not have the ability to carry the large scale data sets in cloud. Its overcome by it in vent the two phase top-down specialization overture. This overture gets input data‘s and split into the small data sets. Small data sets are uniting, then its uses for the anonymization. Here the drawback of proposed system is there is no priority for applying the anonymization on datasets. So, its take more time to anonymize the data sets. So it introduces the scheduling mechanism called Optimized Balanced Scheduling (OBS) to employ the anonymization. Here the OBS means individual dataset have the different sensitive field. It analyses 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) overture to conduct the computation required in TDS in a highly scalable and efficient fashion. The two phases of the overture are based on the two levels of parallelization by MapReduce on cloud. Basically, MapReduce on cloud has two stages of parallelization, i.e., job level and task level. Job level parallelization means that no. of MapReduce jobs can be executed sequentially to make full use of cloud infrastructure resources. Combined with cloud, MapReduce becomes more powerful and flexible 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 can be executed simultaneously over data splits. To achieve high scalability, parallelizing multiple jobs on data separations in the first phase, but the resultant anonymization levels are not same. To obtain finally consistent anonymous data sets, the second phases are necessary to integrate the intermediate results and further anonymize entire data sets. Details are formulated as follows. All intermediate anonymization levels are united into one in the second phase. The uniting of anonymization levels is completed by uniting cuts. Specifically, let in and in be two cuts of an attribute. There exist domain values and it satisfies one of the three conditions is identical to is more general than is more specific than. To assure that the merged intermediate anonymization level never spoil 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 isolation levels, it can merge them in the same way iteratively.

VI. CONCLUSION AND FUTURE WORK

Currently, security in Big data is a challenging research problem. If Integration of MapReduce, a machine for privacy conserving, is designed for the analysing of data would provide better privacy. In the previous 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 conserving data privacy over BigData using MapReduce. Using the two phase top down overture to provide ability to handles the high amount of the large datasets. And here it provides the privacy by effective anonymization overtures.

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 Conservation 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-Conserving Multi-factor Cloud Authentication System Utilizing Big Data‖, IEEE INFOCOM Workshops, pp. 518- 523, 2014.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

[6] Amine Rahmani, Abdelmalek Amine, Reda Mohamed Hamou, ―A Multilayer Evolutionary Homomorphic Coding Overture for Privacy Conserving 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 Overture of Scalable Multidimensional Anonymization for Big Data Privacy Conservation on Cloud‖, Proceedings of 3rdInternational Conference on Cloud and Green Computing, IEEE, pp. 105-112, 2013.
[8] Meiko Jensen and Kiel, ―Challenges of Privacy Protection in Big Data Analytics‖, Proceedings of International Congress on Big Data, IEEE, pp. 235- 238, 2013.
[9] AnjanaGosain, Nikita Chugh, "Privacy Conservation in Big Data", International Journal of Computer Applications (0975 – 8887) Volume 100 – No.17, August 2014.
[10] S. Vennila, J. Priyadarshini, ―Scalable Privacy Conservation in Big Data a Survey‖, Procedia Computer Science 50 (2015) 369-373.
[11] 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.
[12] AntorweepChakravorty,TomaszWlodarczyk, ChunmingRong, ―Privacy Conserving Data Analytics for Smart Homes‖, IEEE Security and Privacy Workshops, pp. 1-5, 2013.
[13] 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.
[14] Linna Li, Michael F. Goodchild and Santa Barbara, ― Is Privacy Still an Issue in the Era of Big Data ―Location disclosure in spatial footprints‖, Proceedings of 21st International conference on Geoinformatics, IEEE, pp.1-4, 2013.
[15] M. Saranya, R and SenthamilSelvi, ―A Survey on Privacy Conservation for AnonymzingData‖ International Journal of Emerging Engineering Research and Technology Volume 3, Issue 1, January 2015, PP 17-21.

BIOGRAPHY

Arundhati Sanjay Reddy is a student of Computer Engineering Department, K J College of Engg. & Management Research , Savitribai Phule Pune University. She received Bachelor of Engg. (BE) degree in 2011 from Swami Ramanand Teerth Marathwada University Nanded, MS India. Now she is pursuing Master's Degree (ME) from KJCOEMR.

Prof. Mininath K. Nighot is a professor in Computer Engineering Department, K J College of Engg. & Management Research , Savitribai Phule Pune University.