Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 50 (2015) 369 – 373

Procedia Computer Science

Scalable Privacy Preservation in Big Data A Survey

Vennila . S[1]
Department of computing science and Engineering
Vellore Institute of Technology
Chennai, India.
s.vennila2014@vit.ac.in

Priyadarshini . J[2]
Department of computing science and Engineering
Vellore Institute of Technology
Chennai, India.
Priyadarshini.J@vit.ac.in

Abstract — Cloud computing provides flexible infrastructure and high storage capacity for BigData applications. The MapReduce framework is most preferable for processing huge volume of unstructured data set in BigData. Increase in data volume leads to flexible and scalable privacy preservation of such dataset over the MapReduce framework is BigData applications. A survey have been taken for the MapReduce framework based big data privacy preservation in Cloud environment. Existing approaches employ local recording anonymization for privacy preserving where data are processed for analysis, mining and sharing. The proposed work focus on Global recording anonymization for preserving data privacy over BigData using MapReduce on Cloud environment.

Keywords— *BigData ; Cloud Computing ; MapReduce; Data Anonymization ; Privacy preservation.*

1. 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 (e.g., healthcare, medical, government services, e-research), 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 refers to hiding identity of sensitive data so that the privacy of an individual is effectively preserved even certain aggregate information can be still exposed to data users for diverse analysis and mining tasks. A variety of privacy models and data anonymization approaches have been proposed and extensively studied [5, 6, 7, 8, 9, 10, 11, 12]. However, applying these traditional approaches to big data anonymization poses scalability and efficiency challenges because of the "3Vs", i.e., Volume,

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)
doi:10.1016/j.procs.2015.04.033

Velocity and Variety. The research on scalability issues of big data anonymization came to the picture [1,2,3,4,9,10] but they lack in some common scenarios.

2. 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 HireSome-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 analyze 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 analyzed 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 recallof 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 via exploiting the idea of the median of medians and the 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. Hence, the future directions for research in this red hot topic are obvious: Antorweep Chakravortyet. al[10] were demonstrated a solution for reliably concealing privacy and ensuring security for analytics of smart home sensor data. The proposed approach maintained the data utility by not transforming the stored data. Rather based on cryptographic techniques, this method replace the personal identifiers of collected sensor data with hashed values before storing them into a identified storage.. The author claimed that the proposed approach is done at design level only.

3. PROPOSED WORK

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.

S.No	Title	Proposed approach	Scalability	Execution price	Availability	Drawback of existing system
1	Proximity-Aware Local-Recoding Anonymization with MapReduce for Scalable Big Data Privacy Preservation in Cloud	Scalable two phase clustering approach	Poor	No	No	preserving approaches tailored to small-scale data sets often fall short when encountering big data, due to their insufficiency
2	HireSome-II: Towards Privacy-Aware Cross-Cloud Service Composition for Big Data Applications	privacy-aware cross-cloud service composition method, named HireSome-II (History record-based Service Optimization method)	Fair	Yes	Yes	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.
3	Engineering Privacy for Big Data Apps with the Unified Modeling Language	Extending vision with privacy Services	No	No	No	extended UML with ribbon icons representing needed big data privacy services.
4	A MapReduce Based Approach of Scalable Multidimensional Anonymization for Big Data Privacy Preservation on Cloud	scalable multidimensional anonymization approach for big data privacy preservation using MapReduce on cloud.	No	No	No	Multidimensional structured data only
5	A Discussion of Privacy Challenges in User Profiling with Big Data Techniques: The EEXCESS Use Case	constructing user profiles with big data techniques and the approaches being considered for preserving user privacy.	No	No	No	User Profile Contents Big Data Techniques for User Profiling Preservation of privacy in big data techniques.
6	Challenges of Privacy Protection in Big Data Analytics	Number of challenges that have to be addressed in order to perform big data analytics in a privacy-compliant way.	No	No	No	anonymization or pseudonymization
7	Privacy Preserving Data Analytics for Smart Homes	an approach to achieve data security & privacy was proposed throughout the complete data lifecycle	No	No	No	Fair
8	Achieving Big Data Privacy via Hybrid Cloud	The time of the AES algorithm; the delay of hybrid-cloud based	No	No	No	Fair
9	Making Big Data, Privacy, and Anonymization work together in the Enterprise: Experiences and Issues	Big Data techniques could yield benefits in the enterprise environment even when working on anonymized data	No	No	No	The effects of removing data on Average Risk Improving Average Risk and Maximum Risk
10	Big data security and privacy issues in healthcare	state-of-the-art security and privacy issues in big data were presented which is applied to healthcare industry.	No	No	No	Fair
11	MACA: A Privacy-Preserving Multi-factor Cloud Authentication System Utilizing Big Data	a privacy-preserving multi-factor authentication system utilizing the features of big data called MACA.	No	No	No	Poor

4. 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 research work proposes Global recording anonymization for preserving data privacy over BigData using MapReduce .

.

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.

[12] 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.