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