Journal of Computer and System Sciences 79 (2013) 542–555

Contents lists available at SciVerse ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

An efficient quasi-identifier index based approach for privacy preservation over incremental data sets on cloud

Xuyun Zhang [a,*], Chang Liu [a], Surya Nepal [b], Jinjun Chen [a]

[a] Faculty of Engineering and Information Technology, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia
[b] Centre for Information & Communication Technologies, Commonwealth Scientific and Industrial Research Organisation, Cnr Vimiera and Pembroke Rodas Marsfield, NSW 2122, Australia

ARTICLE INFO

Article history:
Received 21 March 2012
Received in revised form 18 September 2012
Accepted 8 November 2012
Available online 13 December 2012

Keywords:
Cloud computing
Privacy preservation
Incremental data set
Anonymization
Quasi-identifier index

ABSTRACT

Cloud computing provides massive computation power and storage capacity which enable users to deploy applications without infrastructure investment. Many privacy-sensitive applications like health services are built on cloud for economic benefits and operational convenience. Usually, data sets in these applications are anonymized to ensure data owners' privacy, but the privacy requirements can be potentially violated when new data join over time. Most existing approaches address this problem via re-anonymizing all data sets from scratch after update or via anonymizing the new data incrementally according to the already anonymized data sets. However, privacy preservation over incremental data sets is still challenging in the context of cloud because most data sets are of huge volume and distributed across multiple storage nodes. Existing approaches suffer from poor scalability and inefficiency because they are centralized and access all data frequently when update occurs. In this paper, we propose an efficient quasi-identifier index based approach to ensure privacy preservation and achieve high data utility over incremental and distributed data sets on cloud. Quasi-identifiers, which represent the groups of anonymized data, are indexed for efficiency. An algorithm is designed to fulfil our approach accordingly. Evaluation results demonstrate that with our approach, the efficiency of privacy preservation on large-volume incremental data sets can be improved significantly over existing approaches.

1. Introduction

Technically, cloud computing can be regarded as an ingenious combination of a series of developed or developing ideas and technologies, establishing a pay-as-you-go business model by offering IT services using economies of scale [1–3]. Participants in cloud computing business chains can benefit from this novel business model, as they can save huge IT capital investment by facilitating cloud services such as high storage and computation capabilities, and consequently can concentrate on their own core business [4]. Cloud computing also provides attractive features for science applications in academia [5]. Moreover, since cloud is a multi-tenant environment, it is convenient for cloud users to share data and collaborate with each other. Therefore, many companies or organizations have built up IT systems for their business in cloud computing environments.

However, numerous potential cloud customers are still hesitant to take advantage of cloud computing due to security and privacy concerns [6–10]. Privacy protection is one of the concerned issues in this regard. Currently, several large companies

* Corresponding author.
E-mail addresses: xyzhanggz@gmail.com (X. Zhang), changliu.it@gmail.com (C. Liu), Surya.Nepal@csiro.au (S. Nepal), jinjun.chen@gmail.com (J. Chen).

0022-0000/$ – see front matter
http://dx.doi.org/10.1016/j.jcss.2012.11.008

and hospitals have deployed their health services into cloud, e.g., Microsoft HealthVault [11]. The data sets retained in these cloud applications are highly privacy-sensitive. Once an adversary collects these data sets and menaces the privacy-sensitive information, considerable economic loss or severe social reputation impairment will be caused to corresponding individuals. Usually, these data in cloud will be shared and utilized by multiple users for value-added advantages, rather than just for data storage. Encrypting all the data sets [12–15] is a straightforward and effective privacy protection approach. However, processing effectively and efficiently on encrypted data sets on cloud can be quite a challenging task, because most existing applications run on unencrypted data sets. Recent progress has been made in homomorphic encryption research. Theoretically, computation can be performed on encrypted data sets without decrypting them, but the current techniques are rather expensive and impractical with respect to its efficiency [16,17]. Worse still, encryption still fails to protect individual privacy-sensitive information to legal data users even though it can ensure confidentiality against adversaries. As such, data anonymization techniques like generalization [18] and anatomization [19] have been proposed to preserve privacy when privacy-sensitive data are stored in cloud. Data sets are anonymized to satisfy certain privacy requirements such as k-anonymity [20], or l-diversity [21] before they are shared with data users.

The explosive growth of data sets in cloud applications poses a challenge to existing approaches of privacy preservation over incremental data sets. At present, data sets in many cloud applications often grow incrementally over time as new data are collected and added [22,23]. For instance, a cloud health service will update a huge number of personal health records continuously. Since a large amount of data are generated by special devices or cloud users over time, cloud applications are required to handle incremental data efficiently. To cope with updates of large-volume data sets, a variety of incremental or continuous MapReduce variations have been proposed recently [22–25]. MapReduce is a powerful parallel data processing framework on cloud, offering simple programming model for users [26]. Nevertheless, little attention is paid to privacy issues incurred by incremental privacy-sensitive data in such large-volume data scenarios. When an already anonymized data set is added with new data, two possible effects can take place over the whole data set. One is the violation of privacy requirements, i.e., the newly added data fail to satisfy the anonymity requirement even though they are anonymized according to the current anonymity level. The other is over-anonymization, i.e., the newly added data can be exploited to decrease data distortion by specializing the entire data set to a lower anonymity level, which still satisfies the privacy requirement. Therefore, we should adapt anonymized data sets to achieve both privacy requirement compliance and high data utility when data updates occur. Most existing approaches to address this problem are to re-anonymize the whole updated data set from scratch [27–29], thereby suffering from inefficiency and vulnerability to privacy attacks [30]. Thus, incremental approaches have been proposed accordingly [30–32]. Most of these incremental approaches are centralized while data sets are usually distributed in the context of cloud. A distributed approach has been represented in [32] to preserve k-anonymity over distributed data. However, this approach is inefficient with respect to large-volume data sets on cloud, because it accesses all data when performing data updates. Above all, it is still a challenge to efficiently achieve privacy preservation over distributed and incremental data in the presence of data updates.

In this paper, we propose a novel approach to efficiently achieve privacy preservation over distributed and incremental data sets on cloud. To efficiently update anonymized data sets in the presence of new data, indexing structure of quasi-identifiers is established on anonymized data sets. Quasi-identifiers, which represent the groups of anonymized data, are indexed for efficiency. Moreover, similar data records are placed on the same nodes to reduce communication cost across data storage nodes when anonymized data sets are generalized or specialized to achieve anonymity requirements and high data utility. A commonly used privacy model k-anonymity [20] is employed to measure privacy in our research, i.e., privacy requirements are signified by a threshold k. Further, sub-tree generalization scheme [33] is utilized to accomplish data anonymization. An algorithm is designed to fulfil our approach accordingly. Experimental evaluation on real-world data sets demonstrates that with our approach, the efficiency of privacy preservation over incremental data sets can be improved significantly over existing approaches.

The remainder of this paper is organized as follows. The related work is reviewed in the next section. A motivating example and the problem analysis are given in Section 3. In Section 4 we present our quasi-identifier index based approach and its corresponding algorithm in details. The proposed approach is evaluated in Section 5 by conducting experiments on real-world data sets in our cloud environment. Finally, we conclude this paper and discuss our future work in Section 6.

2. Related work

We briefly present a review on recent research about incremental data processing on cloud, privacy-preserving techniques and privacy preservation over incremental data sets.

Plenty of recent research has investigated the issues of processing incremental data on cloud. Recently, MapReduce has been widely revised from a batch processing framework into a more incremental one to analyze huge-volume incremental data on cloud [34]. Kienzler et al. [24] designed a “stream-as-you-go” approach to access and process on incremental data for data-intensive cloud applications via a steam-based data management architecture. Bhatotia et al. [22] extended the traditional Hadoop framework [35] to a novel one named as Incoop by incorporating several techniques like task partition and memorization-aware schedule. Olston et al. [25] presented a continuous workflow system called Nova on top of Pig/Hadoop through stateful incremental data processing. Li et al. [23] proposed a Hadoop-based platform to support incremental one-pass data analytics by employing hash techniques and a frequent key based technique. However, little attention is paid to