Information Sciences 263 (2014) 198–210

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Dividing secrets to secure data outsourcing

Fatih Emekci [a,*], Ahmed Methwally [b], Divyakant Agrawal [b], Amr El Abbadi [b]

[a] Department of Computer Science, Turgut Ozal University, 06310 Ankara, Turkey
[b] Department of Computer Science, University of California Santa Barbara, Santa Barbara, CA 93106, USA

ARTICLE INFO

Article history:
Received 31 January 2013
Received in revised form 17 July 2013
Accepted 1 October 2013
Available online 16 October 2013

Keywords:
Data outsourcing
Query processing
Data privacy and security

ABSTRACT

Data outsourcing or database as a service is a new paradigm for data management. The third party service provider hosts databases as a service. These parties provide efficient and cheap data management by obviating the need to purchase expensive hardware and software, deal with software upgrades and hire professionals for administrative and maintenance tasks. However, due to recent governmental legislations, competition among companies and database thefts, companies cannot use database service providers directly. They need secure and privacy preserving data management techniques to be able to use them in practice. Since data is remotely stored in a privacy preserving manner, there are efficiency related problems such as poor query response time. We propose a new framework that provides efficient and scalable query response times by reducing the computation and communication costs. Furthermore, the proposed technique uses several service providers to guarantee the availability of the services while detecting the dishonest or faulty service providers without introducing additional overhead on the query response time. The evaluations demonstrate that our data outsourcing framework is scalable and practical.

1. Introduction

Data outsourcing or database as a service is a new paradigm for data management in which a third party service provider hosts database as a service. The service provides data management for its customers and thus obviates the need for the service user to purchase expensive hardware and software, deal with software upgrades and hire professionals for administrative and maintenance tasks. Since using an external database service promises reliable data storage at a low cost by eliminating the need for expensive in-house data-management infrastructure, it is very attractive for companies. However, recent governmental legislations, competition among companies and database thefts have pushed companies to use secure and privacy preserving data management techniques. Using an external database service is a straightforward server–client application in an environment where service providers and clients are honest and clients do not hesitate to share their data with database service providers. However, this is usually not the case and thus the research challenge here is to build a robust and efficient service to manage data in a secure and privacy preserving manner.

Current research has been focused only on how to index and query encrypted data [20,21,9]. Although one of the main problems is querying the encrypted data efficiently, it is not the only problem in data outsourcing. Since thousands of clients per database service provider are expected, the scalability of the proposed techniques and the availability of the services is a very important problem. However, current proposals do not consider this issue and assume a simple scenario consisting of an always available database service provider and a simple service user. Furthermore, they assume both of the parties are honest and trust each other. For example, the service provider may corrupt the data and it would be impossible to recover

* Corresponding author. Tel.: +90 312 5515000.
E-mail address: femekci@turgutozal.edu.tr (F. Emekci).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.10.006

it for the service user. To be able to use external database service providers in real life, there should be a mechanism to recover the data and also to prove that data has been corrupted. *Providing a trust mechanism* to push both database service providers and clients to behave honestly is another important problem.

We propose a new data outsourcing framework providing efficient and scalable query response times. In addition to this, the proposed technique uses multiple service providers to guarantee the availability of the services and to be able to recover from hardware failures. Furthermore, we propose a technique to identify the dishonest or faulty service providers.

Current proposals use encryption to hide the content from service providers [20,9]. However, the computational complexity of encrypting and decrypting data to execute a query increase the query response time. Therefore, this complexity is one of the bottlenecks in current solutions [3]. The proposed solution in this paper uses information theoretically secure techniques similar to Shamir's secret sharing mechanism [29] instead of computationally secure techniques such as encryption. Furthermore, label-based filtration is used to execute range queries [20,22]. However, a data provider reveals some information about the underlying data by labeling a row. Therefore, the computational complexity of our solution is much less than the current proposals using encryption. Therefore, there is a privacy performance tradeoff in these solutions. Our technique does not reveal any information about the content of the data and only the required data is retrieved from the service providers.

In this paper, we use multiple service providers for the fault tolerance. The fault tolerance in this context is the availability of service providers and the ability to recovery from data corruption. Data corruption may happen due to either disk failures or malicious service providers. Our solution deals with both these faults without incurring any additional overhead to the query response time.

The rest of the paper is organized as follows: The model and the types of queries are introduced and also related work is reviewed in Section 2. The basic attempts to solve the problem is discussed in Section 3. Section 4 presents the data distribution technique. The query processing methods for our data distribution technique is studied in Section 5. Section 6 discusses the fault tolerance of the proposed technique. The query response time of the technique is analyzed in Section 7. The last section discusses the future work and concludes the paper.

2. Solution overview and background

In this section, we define the problem and introduce the model. Then, we briefly discuss our solution and finally we review the related work.

2.1. Model and problem formulation

Assume data source D wants to outsource its data to eliminate its database maintenance cost by using the database service provided by database services $DAS_1, \ldots, DAS_n$. D needs to store and access its data remotely without revealing the content of the database to any of the database services. For the sake of this discussion, assume D has a single table $Employees(EID, name, lastname, department, salary)$ in its database and stores *Employees* using the services provided by $DAS_1, \ldots, DAS_n$. After storing *Employees*,D needs to query *Employees* without revealing any information about either the content of the table or queries. Basically, D can pose any of the following queries over time:

1. Exact match queries such as "Retrieve all employees whose name is 'John' ".
2. Range queries such as "Retrieve all employees whose salary is between 10 K and 40 K".
3. Aggregate queries such as MIN/MAX, MEDIAN, SUM and AVERAGE (including aggregate queries over ranges).

There are several proposals addressing exact match queries and range queries [20,9,3], however, these proposals are not complete and do reveal some information about the underlying data (e.g. the range of salaries of employees). In this paper, we will propose a complete approach to execute exact match, range and aggregation queries in a privacy preserving manner. Throughout the paper, we will assume that there are two kinds of attributes in tables namely *numeric attributes* (e.g. salary) and *non-numeric attributes* (e.g. name). The solution first will be presented for numeric attributes and then we will show how to extend it for non-numeric attributes. Throughout this paper, we will develop the work in [20,21] referred to as *data encryption* in parallel with our proposed technique referred to as *secret dividing* so as to show the differences and compare them.

In our solution, data is divided into n shares and each share is stored in a different service provider. When a query is generated at a data source, it is rewritten and the relevant shares are retrieved from the service providers and the query answer is reconstructed at the data source. In order to answer queries, any k of the service providers are needed to be available. n and k are the system parameters and will be discussed later.

2.2. Other related work

Hacigumus et al. [20], Hore et al. [21] and Aggarwal et al. [3] propose using third parties as database service providers. The differences between our work and these work is discussed and compared throughout the paper.