Accepted Manuscript

A multistage protocol for aggregated queries in distributed cloud databases with privacy protection

Andrei Kelarev, Xun Yi, Shahriar Badsha, Xuechao Yang, Leanne Rylands, Jennifer Seberry

PII: S0167-739X(18)30651-4
DOI: https://doi.org/10.1016/j.future.2018.08.017
Reference: FUTURE 4401

To appear in: *Future Generation Computer Systems*

Received date : 23 March 2018
Revised date : 7 August 2018
Accepted date : 8 August 2018

Please cite this article as:, A multistage protocol for aggregated queries in distributed cloud databases with privacy protection, *Future Generation Computer Systems* (2018), https://doi.org/10.1016/j.future.2018.08.017

A multistage protocol for aggregated queries in distributed cloud databases with privacy protection

Andrei Kelarev[a,*], Xun Yi[a], Shahriar Badsha[a], Xuechao Yang[a], Leanne Rylands[b], Jennifer Seberry[c]

[a]*School of Science, RMIT University, GPO Box 2476, Melbourne, VIC 3001, Australia*
[b]*School of Computing, Engineering and Mathematics, Western Sydney University, Locked Bay 1797, Penrith, NSW 2751, Australia*
[c]*School of Computing and Information Technology, University of Wollongong, Northfields Avenue, NSW 2522, Australia*

Abstract

This article is devoted to the novel situation, where a large distributed cloud database is a union of several separate databases belonging to individual database owners who are not allowed to transfer their data for storage in locations different from their already chosen separate cloud service providers. For example, a very large number of medical records may be stored in a distributed cloud database, which is a union of several separate databases from different hospitals, or even from different countries. The owners of the databases may need to provide answers to certain common aggregated queries using all information available without sharing or transferring all data. It is necessary to minimize the communication costs, improve efficiency, and comply with the legal requirements protecting the privacy of confidential data. In this situation, it is impossible to aggregate the whole database in one location, but effective methods for answers to the aggregated queries with privacy protection are required.

To solve this important problem, the present article proposes a Multistage Separate Query Processing (MSQP) protocol employing homomorphic encryption with split keys. We show that our protocol can answer a large class of natural queries of practical significance. The running time of the MSQP protocol is $O(d + \frac{m}{d})$, where d is the number of database owners and m is the total number of records in the whole database. In practice, d is small, m can be very large, and so the running time is $O(m)$. This means that the protocol is very efficient for large databases. It dramatically reduces the communication costs of computation and completely eliminates the need for exchange of confidential data.

We define a new generalized additive homomorphic property and introduce a Multipart ElGamal Cryptosystem (MEC) with split keys, which enjoys this property. MEC is a novel modification of the ElGamal cryptosystem with split keys. This paper presents the results of extensive experiments evaluating the effectiveness of the MSQP protocol employing MEC and comparing it with MSQP employing the ElGamal cryptosystem, for a collection of publicly available medical datasets. The experiments evaluating our protocol on 11 real-life databases and a synthetic database demonstrate that the MSQP protocol employing MEC is more efficient than other options and can be recommended for practical implementations.

Keywords: cloud services, privacy protection, distributed databases, generalized homomorphic property, split keys, multipart ElGamal cryptosystem

1. Introduction

A very large distributed cloud database with privacy protection may need to be created and maintained when several database owners wish to use their separate cloud databases in a collaborative manner by creating a distributed cloud database, but are not allowed to share confidential data contained in their individual databases. For example, a very large number of records can be stored in a distributed database combining data from different hospitals, or even from several different countries (cf. [1, 2]). Databases of this type could be managed by the World Health Organization and can be huge. Different hospitals may be prevented from reallocating the storage of their data without legal consent from the patients. In the case where the database owners are located in different countries, some of these countries may have introduced laws requiring confidential data of patients to be stored in data centers physically located in the same country. Other countries may be considering the introduction of similar privacy laws in the future. An analogous situation may

*Corresponding author
Email addresses: andrei.kelarev@gmail.com (Andrei Kelarev), xun.yi@rmit.edu.au (Xun Yi), shahriar.badsha@rmit.edu.au (Shahriar Badsha), xuechao.yang@rmit.edu.au (Xuechao Yang), l.rylands@westernsydney.edu.au (Leanne Rylands), jennie@uow.edu.au (Jennifer Seberry)