Future Generation Computer Systems 78 (2018) 392–401

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Achieving high performance and privacy-preserving query over encrypted multidimensional big metering data

Rong Jiang [a], Rongxing Lu [b,*], Kim-Kwang Raymond Choo [c,d,e,**]

[a] State Key Laboratory of High Performance Computing and School of Computer, National University of Defense Technology, Changsha, Hunan 410073, China

[b] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

[c] Department of Information Systems and Cyber Security, University of Texas at San Antonio, USA

[d] School of Information Technology and Mathematical Sciences, University of South Australia, Australia

[e] School of Computing, China University of Geosciences, Wuhan, China

HIGHLIGHTS

- Privacy-Preserving Query Over Encrypted Multidimensional Big Metering Data.
- Locality Sensitive Hashing (LSH) based similarity search.
- Encrypted Multidimensional Big Metering Data for Heterogeneous Distributed Systems.

ARTICLE INFO

Article history:
Received 1 February 2016
Received in revised form
7 April 2016
Accepted 5 May 2016
Available online 12 May 2016

Keywords:
Smart grid
High performance
Privacy preservation
Similarity query
Multidimensional big metering data

ABSTRACT

With the proliferation of smart grids, traditional utilities are struggling to handle the increasing amount of metering data. Outsourcing the metering data to heterogeneous distributed systems has the potential to provide efficient data access and processing. In an untrusted heterogeneous distributed system environment, employing data encryption prior to outsourcing can be an effective way to preserve user privacy. However, how to efficiently query encrypted multidimensional metering data stored in an untrusted heterogeneous distributed system environment remains a research challenge. In this paper, we propose a high performance and privacy-preserving query (P2Q) scheme over encrypted multidimensional big metering data to address this challenge. In the proposed scheme, encrypted metering data are stored in the server of an untrusted heterogeneous distributed system environment. A Locality Sensitive Hashing (LSH) based similarity search approach is then used to realize the similarity query. To demonstrate utility of the proposed LSH-based search approach, we implement a prototype using MapReduce for the Hadoop distributed environment. More specifically, for a given query, the proxy server will return K top similar data object identifiers. An enhanced Ciphertext-Policy Attribute-based Encryption (CP-ABE) policy is then used to control access to the search results. Therefore, only the requester with an authorized query attribute can obtain the correct secret keys to retrieve the metering data. We then prove that the P2Q scheme achieves data confidentiality and preserves the data owner's privacy in a semi-trusted cloud. In addition, our evaluations demonstrate that the P2Q scheme can significantly reduce response time and provide high search efficiency without compromising on search quality (i.e. suitable for multidimensional big data search in heterogeneous distributed system, such as cloud storage system).

1. Introduction

With rapid developments in information and communications technologies (ICT), countries such as Australia, Canada, New Zealand, UK and USA are modernizing their aging power system by adopting smart grids (i.e. a heterogeneous distributed system) [1]. Smart grids are characterized by two-way transmissions, high reliability, real-time demand and response, self-healing and higher level of security [2]. In a typical smart grid infrastructure, advanced

* Corresponding author.
** Corresponding author at: School of Information Technology and Mathematical Sciences, University of South Australia, Australia.
E-mail addresses: rxlu@ntu.edu.sg (R. Lu), raymond.choo@fulbrightmail.org (K.-K.R. Choo).

http://dx.doi.org/10.1016/j.future.2016.05.005
0167-739X/

metering infrastructure (AMI) is a key component to ensure grid reliability. AMI also plays a vital role and is closely associated with our daily life [3]. For example, AMI allows the remote reading of meter data, autonomous control of smart appliances (e.g. smart TVs and air conditioning units), and fine-coarse demand response with smart meters and wireless sensor networks [4–6]. Real-time data collected from the smart meters can also improve the reliability of the distributed grids, for example by avoiding line congestion and generation overloads [7,8]. Thus, it is unsurprising that AMI has attracted the attention of stakeholders, such as utility companies, energy markets, and regulators [9,10], as well as security researchers [11]. For example, there are reportedly more than 4.7 million smart meters used for billing and other purposes in Ontario, a province of Canada, alone. Due to the increased computational capacity of newer generation of smart meter and the increasing popularity of smart meters, we are witnessing a significant increase of AMI metering data in recent years. For example, the utility company in Ontario, Canada, reportedly deals with approximately 900 MByte/sec of raw protocol data under a data sending rate of one sample per second and 200 bytes of data per sample. It is also reported that metering data has increased from 10,780 terabytes (TB) in 2010 to over 75,200 TB in 2015 in USA [12].

Metering data, multidimensional in nature, include information such as amount of energy consumed, the time when the energy was consumed, and the purpose of the consumption [13]. An increase in the number of dimensions to be considered will also result in increasing demands for finer grained control and optimization capabilities. However, "How to mine more accurate power consumption patterns or create improved energy usage forecasts from the massive multidimensional data?" remains a challenge, both operationally and academically.

One viable approach to deal with the difficulties in collecting, processing and storing significant amounts of metering data is to outsource such data to a heterogeneous distributed system, such as cloud or fog computing [14]. Cloud servers, for example, offer scalable computing, storage and network resources on demand, which significantly relieve utility companies of the burden in dealing with big (metering) data storage, processing and maintenance. In addition, if metering data are stored in a distributed system environment, users (e.g. customers and other stakeholders) can execute computation and queries 24/7 from anywhere in the world using computational resources available in the distributed system environment.

Similar to any modern day technologies, there are security and privacy risks associated with a distributed system. For example, the untrusted or semi-trusted distributed system, and the hierarchical semi-open network structure in AMI could potentially be targeted by cybercriminals (including state-sponsored actors) seeking to compromise the data for nefarious purpose [15]. It has been reported that over 6 billion USD per year are lost to energy theft in USA [16]. As remarked by Choo [17], the widespread availability of malware, malicious toolkit and other "educational" information online, lowers the technical bar to commit cybercrime. For example, one could easily locate video clips on Youtube and learn how to crack a smart meter to reduce their electricity bill [18]. A customer will also be concerned about the security of their sensitive data and personally identifiable information (PII) (e.g. number of family members and their daily routine that can be inferred from the metering and other collected data). In addition, an unusually low daily power consumption of a household can be an indication that the house occupants are away on holidays. If such data is leaked, the house could be targeted by burglars. Untrusted cloud servers could also share sensitive metering data with third parties for commercial purposes (e.g. targeted advertisements) [19]. Therefore, addressing AMI-related security and privacy issues is a topic of research and policy interest.

In this paper, we seek to protect customers' privacy in smart grids by providing an efficient query mechanism over encrypted multidimensional big metering data in a distributed system. More specifically, we propose a high performance and privacy-preserving query (P2Q) scheme for AMI in smart grids to ensure the confidentiality of multidimensional metering data and customers' privacy. Prior to outsourcing multidimensional metering data to a distributed system such as a cloud, metering data will be encrypted. An enhanced ciphertext-policy attribute-based encryption (CP-ABE) scheme with mixed access structure is adopted to control data acquisition.

We then implement a prototype of the proposed search approach using MapReduce for the Hadoop distribution environment to demonstrate that only requesters with authorized query attributes can perform similarity search and obtain the data via the proposed P2Q scheme. We also analyze the security of the proposed scheme as well as evaluating its performance.

The remainder of this paper is organized as follows. In Section 2, we briefly review related work. In Section 3, we introduce the system model, security model and design goal. Then, in Section 4, we review some preliminaries. In Section 5, we present our proposed scheme, followed by its security and efficiency analysis in Section 6. We conclude the paper in Section 7.

2. Related work

Security and privacy challenges in a distributed smart grid environment are compounded due to the collection, processing and storage of more fine-granular sensitive data, which are then re-used, aggregated, integrated and shared/distributed throughout the entire grid and distributed system. Several privacy preservation schemes for aggregating smart metering data have been proposed in recent years. For example, Lu et al. propose an efficient and privacy-preserving aggregation scheme for smart grid communications [13], which uses a super-increasing sequence to structure multidimensional data and encrypt the structured data using the Paillier cryptosystem technique. The encrypted data can then be aggregated directly in the local gateway without the need for decryption, and the aggregation result of the original data can be obtained at the operation center.

Liang et al. propose a usage-based dynamic pricing (UDP) scheme for smart grid in a community environment, which enables electricity price to correspond to electricity usage in real time [7]. Customer's privacy can be ensured by limiting the disclosure of the individual electricity usage to community gateways. Li et al. use the Merkle hash tree technique to secure smart grid communication [20], and more specifically against replay, message injection, message analysis, and message modification attacks.

A number of approaches have also been proposed to perform privacy-preserving data aggregation in AMI [21], such as homomorphic encryption and secret sharing [22], masking and brute forcing [23], modified homomorphic encryption [24], masking and differential privacy [25]. However, query schemes over encrypted data in smart grids, critical for a user's metering data audit, have been under-studied in the literature. For example, Wen et al. propose a novel privacy-preserving range query (PaRQ) scheme over encrypted metering data to address privacy issues in financial auditing for smart grid [3]. The PaRQ scheme constructs a range query predicate based on hidden vector encryption, and provides an efficient range search while preserving data confidentiality and query privacy. However, the scheme is not able to provide similarity search over encrypted multidimensional data.

Similarity search in high-dimensional spaces has also attracted the attention of various researchers, due to its importance in databases, data mining, and search engines particularly in a cloud storage environment. Locality Sensitive Hashing (LSH) [26]