Secure and Privacy-Preserving Data Services in the Cloud: A Data Centric View *

Divyakant Agrawal Amr El Abbadi Shiyuan Wang
Department of Computer Science, UC Santa Barbara
Santa Barbara, CA 93106-5110, USA
{agrawal, amr, sywang}@cs.ucsb.edu

ABSTRACT

Cloud computing becomes a successful paradigm for data computing and storage. Increasing concerns about data security and privacy in the cloud, however, have emerged. Ensuring security and privacy for data management and query processing in the cloud is critical for better and broader uses of the cloud. This tutorial covers some common cloud security and privacy threats and the relevant research, while focusing on the works that protect data confidentiality and query access privacy for sensitive data being stored and queried in the cloud. We provide a comprehensive study of state-of-the-art schemes and techniques for protecting data confidentiality and access privacy, which make different tradeoffs in the multidimensional space of security, privacy, functionality and performance.

1. INTRODUCTION

Cloud computing has emerged as a successful paradigm that considerably simplifies the deployment of computing and storage infrastructures of both large and small enterprises. Increasing concerns about data security and privacy in the cloud, however, have emerged, as vulnerabilities were found in cloud service providers' sites [15], and user data leakage incidents were reported for a number of cloud based application services. Ensuring security and privacy for data management and query processing in the cloud is therefore critical for better and broader uses of the cloud.

Nevertheless, providing such secure and privacy-preserving data services is very challenging, as security problems can arise in multiple levels of the data services, and security and privacy protection may impede functionality and performance of the data services. This tutorial aims to cover some common cloud security and privacy threats and the relevant research, while focusing on the works that protect data confidentiality and query access privacy for sensitive data being stored and queried in the cloud. We provide a comprehensive study of state-of-the-art schemes and techniques for protecting data confidentiality and access privacy, which make different tradeoffs in the multidimensional space of security, privacy, functionality and performance. We also identify their limitations and further discuss future research directions in cloud data security and privacy.

*This work is supported by NSF Grants CNS-1053594 and III-1018637.

Proceedings of the VLDB Endowment, Vol. 5, No. 12

2. TUTORIAL OUTLINE

2.1 Security and Privacy Threats

We start our tutorial by presenting a general overview of various security and privacy threats that could arise in the context of data services deployed in the cloud. We consider the cloud service providers and any unauthorized parties that can monitor, affect or control the cloud activities as *adversaries*. We identify the desirable features for ensuring a secure and privacy-preserving data service in the cloud. We focus on the critical features of ensuring *data confidentiality* and *access privacy*, as they usually conflict with the normal functioning and performance of data services in the cloud, and thus give rise to numerous research challenges.

2.2 Data Confidentiality

In this section of the tutorial, we discuss schemes and techniques for ensuring data confidentiality while allowing data management and query processing on the protected data in the cloud. To protect the confidentiality of sensitive private data stored in the cloud, encryption is a standard technique. Encrypting the data however makes it difficult for the cloud to process queries on the data for users, thus various techniques have been proposed for querying on encrypted data. Alternatively, we can explore trusted computing instead of encryption and querying on encrypted data. In the following, we delve into the details of various approaches that ensure data confidentiality: (i) data encryption and querying on encrypted data, and (ii) trusted computing.

Data Encryption and Querying on Encrypted Data. This part considers data are encrypted in the cloud and the data should not be disclosed to adversaries during query processing, while the adversaries could launch statistical analysis and inference attacks to infer the data contents. We provide a survey of the techniques for processing various database queries such as range queries and aggregation queries on encrypted data. Processing range queries requires the ability to compare a ciphertext data value with the encrypted query range boundary values. This can be achieved by providing the cloud rough information about the data [1, 8, 9, 14] or building obfuscated index structures in the cloud [4, 18]. Processing aggregation queries is usually

achieved by using a special encryption scheme called *homomorphic encryption* that allows addition and multiplication on ciphertexts without the need for decryption [6, 13]. We then briefly introduce *fully homomorphic encryption* that allows arbitrary computation on ciphertexts [7].

Trusted Computing. An alternative to data encryption and querying on encrypted data is to keep the plaintext data in a secure trusted container in the cloud [2]. We present this idea of *trusted computing* as the last part of data confidentiality. Finally we compare data encryption and query on encrypted data with trusted computing from the three angles of security, performance and databases.

2.3 Access Privacy

When the data is being queried, queries may reveal partial information about the data. Hence, ensuring access privacy is important. In this section of the tutorial, we survey the most representative cryptographic protocols for protecting access privacy in general: *Private Information Retrieval* [3], a special memory structure that obfuscates query access patterns over encrypted data, *Oblivious RAM* [12], and practical alternative techniques for protecting access privacy.

Private Information Retrieval. Private Information Retrieval (PIR) solves the problem of privately retrieving a data item from a remote database server without revealing to the server which item is retrieved [3]. There are PIR solutions that only use one server [10] and solutions that rely on multiple servers [3]. Earlier single server PIR solutions have been criticized for their significant computation overhead [16], while a more recent study based on recent developments of PIR demonstrates that PIR can be practical [11]. Our goal here is to understand the basic rational of PIR protocols and discuss the practicality of these protocols in terms of computation and communication costs as well as from the cloud service point of view.

Oblivious RAM. One way to make PIR more practical, as proposed in [19], is to employ an oblivious RAM [12] on the cloud server. The basic idea of oblivious RAM is to shuffle and re-sort data items in the RAM during data accesses.

Practical Alternative Techniques. Some alternative techniques have been proposed to achieve practical access privacy, such as covered search and index shuffling for protecting accesses to encrypted index [5], and hybrid approaches that apply PIR operations on selected partial data [17]. Our goal in this part is to learn the basic ideas of these techniques and understand the privacy performance tradeoffs they made.

3. GOALS OF THE TUTORIAL

3.1 Learning Outcomes

- Overview of security and privacy threats in the cloud.
- State-of-the-art in data confidentiality and access privacy protection for data management and query processing services in the cloud.
- Understanding the tradeoffs of query performance, data confidentiality and access privacy of various protection techniques.
- Overview of other studies in cloud security.

3.2 Intended Audience

This tutorial is intended to benefit researchers and system designers in security, privacy and the cloud. An understanding of current research and systems is essential for designing new protection techniques and building secure data services.

4. BIOGRAPHICAL SKETCHES

Divyakant Agrawal is a Professor of Computer Science at University of California, Santa Barbara. His research expertise is in the areas of database systems, distributed computing, data warehousing, and large-scale information systems.

Amr El Abbadi is a Professor of Computer Science at University of California, Santa Barbara. His research interests lie in the area of scalable database and distributed systems.

Shiyuan Wang is a PhD Candidate in Computer Science Department at University of California Santa Barbara. Her research interests are data security and privacy.

5. REFERENCES

[1] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu. Order preserving encryption for numeric data. In *SIGMOD*, pages 563–574, 2004.

[2] S. Bajaj and R. Sion. TrustedDB: a trusted hardware based database with privacy and data confidentiality. In *SIGMOD*, pages 205–216, 2011.

[3] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan. Private information retrieval. *Journal of ACM*, 45(6):965–981, 1998.

[4] E. Damiani, S. D. C. di Vimercati, S. Jajodia, S. Paraboschi, and P. Samarati. Balancing confidentiality and efficiency in untrusted relational DBMSs. In *CCS*, pages 93–102, 2003.

[5] S. De Capitani di Vimercati, S. Foresti, S. Paraboschi, G. Pelosi, and P. Samarati. Efficient and private access to outsourced data. In *ICDCS*, pages 710–719, 2011.

[6] T. Ge and S. B. Zdonik. Answering aggregation queries in a secure system model. In *VLDB*, pages 519–530, 2007.

[7] C. Gentry. Computing arbitrary functions of encrypted data. *Communication of ACM*, 53:97–105, 2010.

[8] H. Hacigumus, B. R. Iyer, C. Li, and S. Mehrotra. Executing SQL over encrypted data in the database service provider model. In *SIGMOD*, pages 216–227, 2002.

[9] B. Hore, S. Mehrotra, and G. Tsudik. A privacy-preserving index for range queries. In *VLDB*, pages 720–731, 2004.

[10] E. Kushilevitz and R. Ostrovsky. Replication is not needed: Single database, computationally-private information retrieval. In *FOCS*, pages 364–373, 1997.

[11] F. G. Olumofin and I. Goldberg. Revisiting the computational practicality of private information retrieval. In *Financial Cryptography*, pages 158–172, 2011.

[12] R. Ostrovsky. Efficient computation on oblivious RAMs. In *STOC*, pages 514–523, 1990.

[13] P. Paillier. Public-key cryptosystems based on composite degree residuosity classes. In *EUROCRYPT*, pages 223–238, 1999.

[14] R. A. Popa, C. M. S. Redfield, N. Zeldovich, and H. Balakrishnan. CryptDB: protecting confidentiality with encrypted query processing. In *SOSP*, pages 85–100, 2011.

[15] T. Ristenpart, E. Tromer, H. Shacham, and S. Savage. Hey, you, get off of my cloud: exploring information leakage in third-party compute clouds. In *CCS*, pages 199–212, 2009.

[16] R. Sion and B. Carbunar. On the computational practicality of private information retrieval. In *NDSS Symposium*, 2007.

[17] S. Wang, D. Agrawal, and A. E. Abbadi. Generalizing pir for practical private retrieval of public data. In *DBSec*, pages 1–16, 2010.

[18] S. Wang, D. Agrawal, and A. El Abbadi. A comprehensive framework for secure query processing on relational data in the cloud. In *SDM*, pages 52–69, 2011.

[19] P. Williams and R. Sion. Usable private information retrieval. In *NDSS Symposium*, 2008.