Secure Data Outsourcing

Radu Sion
Computer Science
Stony Brook University
Stony Brook, NY 11794
sion@cs.stonybrook.edu
Phone: (631) 632 - 1672
Fax: (631) 632 - 1690

Summary. The networked and increasingly ubiquitous nature of today's data management services mandates assurances to detect and deter malicious or faulty behavior. This is particularly relevant for outsourced data frameworks in which clients place data management with specialized service providers. Clients are reluctant to place sensitive data under the control of a foreign party without assurances of confidentiality. Additionally, once outsourced, privacy and data access correctness (data integrity and query completeness) become paramount. Today's solutions are fundamentally insecure and vulnerable to illicit behavior, because they do not handle these dimensions.

In this tutorial we will explore how to design and build robust, efficient, and scalable data outsourcing mechanisms providing strong security assurances of (1) *correctness*, (2) *confidentiality*, and (3) data access *privacy*.

There exists a strong relationship between such assurances; for example, the lack of access pattern privacy usually allows for statistical attacks compromising data confidentiality. Confidentiality can be achieved by data encryption. However, to be practical, outsourced data services should allow expressive client queries (e.g., relational joins with arbitrary predicates) without compromising confidentiality. This is a hard problem because decryption keys cannot be directly provided to potentially untrusted servers. Moreover, if the remote server cannot be fully trusted, protocol correctness become essential. Therefore, solutions that do not address all three dimensions are incomplete and insecure.

We will discuss query mechanisms targeting outsourced relational data that (i) ensure queries have been executed with integrity and completeness over their respective target data sets, (ii) allow queries to be executed with confidentiality over encrypted data, (iii) guarantee the privacy of client queries and data access patterns. We will explore protocols that adapt to the existence of *trusted hardware* — so critical functionality can be delegated securely from clients to servers.

Audience: extremely broad. The intended audience is extremely broad, to include researchers in all areas of information processing that involve storage and out-sourcing of valuable data. The total cost of ownership of data management infra-structure is 5–10 times greater than the hardware costs, and more data is produced and lives digitally every day. In the coming years, secure, robust, and efficient outsourced data management will be demanded by users. Understanding the challenges and the available solutions in these areas is essential.

Proposed Length: 3 hours. This span optimally accommodates the material, averaging 1 hour for each of the discussed dimensions (correctness, confidentiality, and access privacy). I have given shorter versions of this tutorial before and I am extremely confident this is the exact amount of time required to fit the material. The audience is expected to gain a solid understanding of secure data outsourcing and its main research and implementation issues.

Pre-requisites: none. The tutorial is designed to only require broad knowledge of computer science. A general introduction to data security will be included as part of the tutorial.

Previous instances: IIT Delhi, December 2006 A shorter version of this tutorial has been invited to IIT Delhi (as part of COMAD) in December 2006. The VLDB version significantly differs in both length and scope. The presenter will now focus on the applied (existing and future proposed) solutions dimension (as opposed to discussing mainly challenges).

Biography of Speaker. Radu Sion is an Assistant Professor of Computer Sciences in Stony Brook University and the director of the Network Security and Applied Cryptography Laboratory. His research focuses on data security and information assurance mechanisms. Collaborators and funding partners include Motorola Labs, IBM Research, the Center of Excellence in Wireless and Information Technology CEWIT, the Stony Brook Office for the Vice-President for Research and the National Science Foundation. Dr. Sion is serving on the organizing committee of numerous data management and information security conferences, such as SIGMOD, ICDE, ICDCS, CCS, Financial Cryptography, USENIX Security a.o.

Overview. Today, sensitive data is being managed on remote servers maintained by third party outsourcing vendors. This is because the total cost of data management is 5–10 times higher than the initial acquisition costs [29]. In such an outsourced "database as a service" [31] model, *clients* outsource data management to a "database service provider" that provides online access mechanisms for querying and managing the hosted data sets.

This is advantageous and significantly more affordable for parties with limited abilities to manage large in-house data centers of potentially large resource footprints. By comparison, database service providers [1–6,6–9,11–15] – ranging from corporate-level services such as the IBM Data Center Outsourcing Services to personal level database hosting – have the advantage of expertize consolidation. More-over they are likely to be able to offer the service much cheaper, with increased service availability (e.g. uptime) guarantees.

Notwithstanding these clear advantages, a data outsourcing paradigm faces significant challenges to widespread adoption, especially in an online, untrusted environment. Current privacy guarantees of such services are at best declarative and often subject customers to unreasonable fine-print clauses—e.g., allowing the server operator (and thus malicious attackers gaining access to its systems) to use customer behavior and content for commercial, profiling, or governmental surveillance purposes [27]. Clients are naturally reluctant to place sensitive data under the control of a foreign party without strong security assurances of *correctness* [28, 33, 41, 45, 47], *confidentiality* [10, 16], and data access *privacy* [22–26, 36, 37, 40, 42, 48, 49]. These assurances are essential for data outsourcing to become a sound and truly viable alternative to in-house data management. However, developing assurance mechanisms in such frameworks is challenging because the data is placed under the authority of an external party whose honest behavior is not guaranteed but rather needs to be ensured by this very solution.

In this tutorial, we will explore the challenges of designing and implementing robust, efficient, and scalable relational data outsourcing mechanisms, with strong security assurances of *correctness*, *confidentiality*, and data access *privacy*. This is important because today's outsourced data services are fundamentally insecure and vulnerable to illicit behavior, because they do not handle all three dimensions consistently and there exists a strong relationship between such assurances: e.g., the lack of access pattern privacy usually allows for statistical attacks compromising data confidentiality. Even if privacy and confidentiality are in place, to be practical, outsourced data services should allow sufficiently expressive client queries (e.g., relational operators such as JOINs with arbitrary predicates) without compromising confidentiality. This is a hard problem because in most cases decryption keys cannot be directly provided to potentially untrusted database servers. Moreover, result completeness and data integrity (i.e., correctness) become essential. Therefore, solutions that do not address these dimensions are incomplete and insecure.

We will explore designs for outsourced relational data query mechanisms that (i) ensure queries have been executed with *integrity and completeness* over their respective target data sets, (ii) allow queries to be executed with *confidentiality* over encrypted data, (iii) guarantee the *privacy* of client queries and data access patterns. We will discuss protocols that adapt to the existence of *trusted hardware* — so critical functionality can be delegated securely from clients to servers and increased assurance levels can be achieved more efficiently. Moreover, it is important to design for scalability to large data sets and high query throughputs. We discuss implementation issues in achieving the above three security assurances:

Correctness. Clients should be able to verify the integrity and completeness of any results the server returns. For example, when executing a JOIN query, they should be able to verify that the server returned *all* matching tuples.

Confidentiality. The data being stored on the server should not be decipherable either during transit between the client and the server, or at the server side, even in the case when the server is malicious.

Access Privacy. An intruder or a malicious server should not be able to perform statistical attacks by exploiting query patterns. For example, it should not be able to compromise data confidentiality by correlating known public information with frequently queried data items.

Authentication/Authorization. We note that client *authentication* and *authorization*, two important but orthogonal security dimensions, are extensively addressed elsewhere [17–21, 30, 32, 34, 35, 38, 39, 43, 44, 46, 50]; therefore they and are not the main focus here. The assurances discussed here naturally complement these dimensions in providing increased end-to-end security.

References

[1] Activehost.com Internet Services. Online at http://www.activehost.com.

[2] Adhost.com MySQL Hosting. Online at http://www.adhost.com.

[3] Alentus.com Database Hosting. Online at http://www.alentus.com.

[4] Datapipe.com Managed Hosting Services. Online at http://www.datapipe.com.

[5] Discountasp.net Microsoft SQL Hosting. Online at http://www.discountasp.net.

[6] Gate.com Database Hosting Services. Online at http://www.gate.com.

[7] Hostchart.com Web Hosting Resource Center. Online at http://www.hostchart.com.

[8] Hostdepartment.com MySQL Database Hosting. Online at http://www.hostdepartment.com/mysqlwebhosting/.

[9] IBM Data Center Outsourcing Services. Online at http://www-1.ibm.com/services/.

[10] IBM Data Encryption for DB2. Online at http://www.ibm.com/software/data/db2.

[11] Inetu.net Managed Database Hosting. Online at http://www.inetu.net.

[12] Mercurytechnology.com Managed Services for Oracle Systems. Online at http://www.mercurytechnology.com.

[13] Neospire.net Managed Hosting for Corporate E-business. Online at http://www.neospire.net.

[14] Netnation.com Microsoft SQL Hosting. Online at http://www.netnation.com.

[15] Opendb.com Web Database Hosting. Online at http://www.opendb.com.

[16] Oracle: Database Encryption in Oracle 10g. Online at http://www.oracle.com/database.

[17] Martin Abadi, Michael Burrows, Butler Lampson, and Gordon Plotkin. A calculus for access control in distributed systems. *ACM Trans. Program. Lang. Syst.*, 15(4):706–734, 1993.

[18] Steven M. Bellovin. Spamming, phishing, authentication, and privacy. *Communications of the ACM*, 47(12):144, 2004.

[19] Elisa Bertino, Sushil Jajodia, and Pierangela Samarati. A flexible authorization mechanism for relational data management systems. *ACM Transactions on Information Systems*, 17(2), 1999.

[20] Ray Bird, Inder Gopal, Amir Herzberg, Phil Janson, Shay Kutten, Refik Molva, and Moti Yung. The kryptoknight family of light-weight protocols for authentication and key distribution. *IEEE/ACM Trans. Netw.*, 3(1):31–41, 1995.

[21] M. Burrows, M. Abadi, and R. Needham. A logic of authentication. In *SOSP '89: Proceedings of the twelfth ACM symposium on Operating systems principles*, pages 1–13, New York, NY, USA, 1989. ACM Press.

[22] C. Cachin, S. Micali, and M. Stadler. Computationally private information retrieval with polylog communication. In *Proceedings of EUROCRYPT*, 1999.

[23] C. Cachin, S. Micali, and M. Stadler. Private Information Retrieval with Polylogarithmic Communication. In *Proceedings of Eurocrypt*, pages 402–414. Springer-Verlag, 1999.

[24] Y. Chang. Single-Database Private Information Retrieval with Logarithmic Communication. In *Proceedings of the 9th Australasian Conference on Information Security and Privacy ACISP*. Springer-Verlag, 2004.

[25] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan. Private information retrieval. In *IEEE Symposium on Foundations of Computer Science*, pages 41–50, 1995.

[26] Benny Chor, Eyal Kushilevitz, Oded Goldreich, and Madhu Sudan. Private information retrieval. *J. ACM*, 45(6):965–981, 1998.

[27] CNN. Feds seek Google records in porn probe. Online at http://www.cnn.com, January 2006.

[28] Premkumar T. Devanbu, Michael Gertz, Chip Martel, and Stuart G. Stubblebine. Authentic third-party data publication. In *IFIP Workshop on Database Security*, pages 101–112, 2000.

[29] Gartner, Inc. Server Storage and RAID Worldwide. Technical report, Gartner Group/Dataquest, 1999. www.gartner.com.

[30] Li Gong. Efficient network authentication protocols: lower bounds and optimal implementations. *Distrib. Comput.*, 9(3):131–145, 1995.

[31] H. Hacigumus, B. R. Iyer, and S. Mehrotra. Providing database as a service. In *IEEE International Conference on Data Engineering (ICDE)*, 2002.

[32] E. Hildebrandt and G. Saake. User Authentication in Multidatabase Systems. In R. R. Wagner, editor, *Proceedings of the Ninth International Workshop on*

Database and Expert Systems Applications, August 26–28, 1998, Vienna, Austria, pages 281–286, Los Alamitos, CA, 1998. IEEE Computer Society Press.

[33] HweeHwa Pang and Arpit Jain and Krithi Ramamritham and Kian-Lee Tan. Verifying Completeness of Relational Query Results in Data Publishing. In *Proceedings of ACM SIGMOD*, 2005.

[34] S. Jajodia, P. Samarati, and V. S. Subrahmanian. A Logical Language for Expressing Authorizations. In *IEEE Symposium on Security and Privacy*, pages 31–42, Oakland, CA, May 04-07 1997. IEEE Press.

[35] S. Jajodia, P. Samarati, and V. S. Subrahmanian. A logical language for expressing authorizations. In *IEEE Symposium on Security and Privacy. Oakland, CA*, pages 31–42, 1997.

[36] E. Kushilevitz and R. Ostrovsky. Replication is not needed: single database, computationally-private information retrieval. In *Proceedings of FOCS*. IEEE Computer Society, 1997.

[37] E. Kushilevitz and R. Ostrovsky. One-way trapdoor permutations are sufficient for non-trivial single-server private information retrieval. In *Proceedings of EUROCRYPT*, 2000.

[38] Butler Lampson, Mart&#237;n Abadi, Michael Burrows, and Edward Wobber. Authentication in distributed systems: theory and practice. *ACM Trans. Comput. Syst.*, 10(4):265–310, 1992.

[39] Li, Feigenbaum, and Grosof. A logic-based knowledge representation for authorization with delegation. In *PCSFW: Proceedings of the 12th Computer Security Foundations Workshop*, 1999.

[40] H. Lipmaa. An oblivious transfer protocol with log-squared communication. Cryptology ePrint Archive, 2004.

[41] Maithili Narasimha and Gene Tsudik. Authentication of Outsourced Databases using Signature Aggregation and Chaining. In *Proceedings of DASFAA*, 2006.

[42] E. Mann. Private access to distributed information. Master's thesis, Technion - Israel Institute of Technology, 1998.

[43] Fabian Monrose and Aviel D. Rubin. Authentication via keystroke dynamics. In *ACM Conference on Computer and Communications Security*, pages 48–56, 1997.

[44] Fabian Monrose and Aviel D. Rubin. Keystroke dynamics as a biometric for authentication. *Future Generation Computer Systems*, 16(4):351–359, 2000.

[45] E. Mykletun, M. Narasimha, and G. Tsudik. Authentication and integrity in outsourced databases. In *ISOC Symposium on Network and Distributed Systems Security NDSS*, 2004.

[46] Roger M. Needham and Michael D. Schroeder. Using encryption for authentication in large networks of computers. *Commun. ACM*, 21(12):993–999, 1978.

[47] Radu Sion. Query execution assurance for outsourced databases. In *Proceedings of the Very Large Databases Conference VLDB*, 2005.

[48] Radu Sion and Bogdan Carbunar. On the Practicality of Private Information Retrieval. In *Proceedings of the Network and Distributed Systems Security Symposium*, 2007. Stony Brook Network Security and Applied Cryptography Lab Tech Report 2006-06.

[49] J. Stern. A new and efficient all-or-nothing disclosure of secrets protocol. In *Proceedings of Asia Crypt*, pages 357–371, 1998.

[50] Thomas Y. C. Woo and Simon S. Lam. Authentication for distributed systems. *Computer*, 25(1):39–52, 1992.