Chapter 15
Using Data Masking for Balancing Security and Performance in Data Warehousing

Ricardo Jorge Santos
CISUC – FCTUC – University of Coimbra, Portugal

Jorge Bernardino
CISUC – ISEC – Polytechnic Institute of Coimbra, Portugal

Marco Vieira
CISUC – FCTUC – University of Coimbra, Portugal

ABSTRACT

Data Warehouses (DWs) are the core of sensitive business information, which makes them an appealing target. Encryption solutions are accepted as the best way to ensure strong security in data confidentiality while keeping high database performance. However, this work shows that they introduce massive storage space and performance overheads to a magnitude that makes them unfeasible for DWs. This work proposes a data masking technique for protecting sensitive business data in DWs which balances security strength with database performance, using a formula based on the mathematical modular operator and simple arithmetic operations. The proposed solution provides apparent randomness in the generation and distribution of the masked values, while introducing small storage space and query execution time overheads. It also enables a false data injection method for misleading attackers and increasing the overall security strength. It can be easily implemented in any DataBase Management System (DBMS) and transparently used, without changes to application source code. Experimental evaluations using a real-world DW and TPC-H decision support benchmark implemented in leading DBMS Oracle 11g and Microsoft SQL Server 2008 demonstrate its overall effectiveness. Results show the substantial savings of its implementation costs when compared with state of the art encryption solutions provided by those DBMS and that it outperforms those solutions in both data querying and insertion of new data.

DOI: 10.4018/978-1-4666-2518-1.ch015

INTRODUCTION

Data confidentiality focuses on protecting data from unauthorized disclosure. Currently, data is a major asset for any enterprise, not only for knowing the past, but also for aiding today's business or predicting future trends (Baer, 2004; Kobielus, 2009). Given its decision support nature, the data in Data Warehouses (DWs) translates into business knowledge, providing invaluable decision making information for adding business value. Consequently, DWs are the core of the enterprise's sensitive data. Unfortunately, this makes them a main target for both inside and outside attackers (Yuhanna, 2009). Consequently, several studies have demonstrated that efficiently securing sensitive data has become an imperative concern in many enterprises (McKendrick, 2009; Yuhanna, 2009).

To protect information in databases, data masking actions and encryption techniques are widely used. Data masking routines are mainly simpler than encryption routines, but provide lower security strength. Moreover, data masking routines provided by most commercial tools typically change data in an irreversible manner, *i.e.*, after masking data it not possible to subsequently retrieve the original true values, making them useless for real live DW databases. This has made masking solutions the main choice for protecting published data or production data, instead of real-live databases (Bertino & Sandhu, 2005; Huey, 2008; Natan, 2005; Oracle, 2010a; Ravikumar et al., 2011; Gartner, 2009).

Published research and best practice guides state that encryption is the best method to protect sensitive data at the database level while maintaining high database performance (Agrawal et al., 2004; Ge & Zdonik, 2007; Huey, 2008; Natan, 2005; Oracle, 2010a, 2010b; Procopiuc & Srivastava, 2011; Vimercati et al., 2007; Hacigumus et al., 2004). However, since DWs are usually huge in size, with millions or billions of rows in their fact tables, and user queries are typically *ad hoc* and access large amounts of data, encryption and decryption overhead is a major concern (Agrawal et al., 2004). General encryption algorithms are mainly built taking under consideration desirable security strength. Although efficient in their security purpose, most encryption solutions introduce several key costs from the DW perspective:

- Large processing time/resources for encrypting sensitive data, given routine or hardware access in very large databases such as those in DWs;
- Extra storage space of encrypted data. Since DWs usually have many millions or billions of rows, even a small modification of any column size to accommodate encrypted output introduces large storage space overheads; and
- Overhead query response time and allocated resources for decrypting data to process queries. Given the huge amount of data typically accessed by DW queries, this is probably the most significant drawback for using encryption in DWs.

In this work, we present the commonly available encryption techniques for databases, thoroughly discussing the issues involving the use of these techniques, in what concerns database performance and requirements, from the data warehousing perspective. As demonstrated throughout this work, the introduced overheads imply that the standard encryption algorithms currently provided by DBMS dramatically degrade database performance, which is a critical issue in data warehousing. This ultimately jeopardizes their applicability in DWs. Consequently, developing a data masking/encryption strategy for DWs must balance between the requirements for security and desire for high performance (Ge & Zdonik, 2007; Mattson, 2004; Nadeem & Javed, 2005; Vieira & Madeira, 2005).