

Enterprise Data Management Solutions
October 2008



IBM **Information Management** software

Closing the data privacy gap for SAP Applications

Contents

2 Executive summary

3 Why is data privacy a high priority?

**4 Closing the data privacy gap in
non-production environments**

**5 Selecting a comprehensive data
privacy solution**

**7 Meeting data privacy challenges
with IBM Optim**

8 Proven data masking techniques

**14 Data privacy best practices
summary**

Executive summary

This white paper explains why protecting confidential information and ensuring privacy have become high priorities. News headlines about the increasing frequency of stolen information and identity theft have focused awareness on privacy breaches and their consequences. In responding to these issues, data privacy regulations have been enacted around the world. Although the specifics of these regulations may differ, failure to ensure data privacy compliance can result in millions of dollars in financial penalties and jail time. Companies also risk losing customer loyalty and destroying brand equity. The impact is serious enough to put a company out of business.

Companies rely on critical SAP® applications to support daily business operations, so it is essential to ensure privacy and protect application data no matter where it resides. However, the same methods that protect data in production environments may not meet the unique requirements for non-production (development, testing and training) environments. How can IT organizations protect personal data, including employee and customer information, as well as corporate confidential data and intellectual property? Industry analysts recommend “de-identifying” or masking data as a best practice for protecting privacy. But, what are some of the requirements for selecting a data privacy solution?

The ideal data privacy solution must provide the necessary data masking techniques to satisfy both the simplest and most complex privacy requirements. Masking techniques must also produce results that reflect the application logic and preserve the integrity of the data. To help support your data privacy compliance requirements, this white paper describes some of the comprehensive data masking techniques available with the IBM® Optim™ Data Privacy Solution for SAP® Applications:

- Application-aware masking capabilities help ensure that masked data, like names and street addresses, resembles the look and feel of the original information.

- Context-aware, prepackaged data masking routines make it easy to de-identify data elements, such as Social Security numbers, payroll data and e-mail addresses.
- Persistent masking capabilities propagate masked replacement values consistently across applications, databases, operating systems and hardware platforms.

With Optim, companies can de-identify data in a way that is valid for use in development, testing and training environments, while protecting data privacy.

Why is data privacy a high priority?

The information explosion has made access to public and private information a part of everyday life. SAP applications typically collect this information for legitimate purposes; however, given the interconnected nature of the Internet and information systems, personal data is often subject to theft and misuse.

Identity theft, privacy violations and fraudulent access to sensitive information continue to make news headlines. As more companies recognize the need for data privacy, and customers demand increased protection, government entities are passing increasingly stringent laws and regulations to ensure that safety. Every company must be responsible for protecting confidential employee information, corporate business intelligence and sensitive customer data to comply with information governance regulations and to gain the trust of customers and business partners.

Data privacy compliance affects your business. Protecting data privacy is a critical business initiative. More than 70 percent of data breaches actually result from internal weaknesses.¹ Examples range from employees, who misuse payment card numbers and other sensitive information, to those who save confidential data on laptops that are subsequently stolen. In addition, outsourcing application data to off-shore processing environments makes it difficult to control access to otherwise unsecured sensitive data and to comply with Safe Harbor directives.

More than 70 percent of data breaches actually result from internal weaknesses.

The stakes are high. Corporations and their officers face fines that can reach \$500,000 per incident and can include jail time. Hard penalties are only one example of how organizations can be harmed. Other negative impacts include erosion in share price caused by investor concern and negative publicity resulting from a data breach. Irreparable brand damage identifies a company as one that cannot be trusted. Companies that do not take steps to protect confidential information risk not only losing customers and revenue, but also risk going out of business. If losing customers does not have a financial impact on your business, consider that investigating a breach can cost millions.

Closing the data privacy gap in non-production environments

Depending on the industry, operations and types of applications, many production and non-production instances will process confidential information. The challenge is to provide the appropriate protection, while meeting business needs and ensuring that data is managed on a “need-to-know” basis.

Most companies manage multiple production instances of their SAP applications. For example, a company that has implemented SAP HCM (Human Capital Management) may deploy one application instance to support its North American operations, a second for EMEA and a third for APAC. To support application development, testing, training, backup and other activities, a site may manage anywhere from 3 to 30 copies of each instance, containing an exact replica of the confidential data from the source system.

Companies can protect private information in their SAP production transaction processing environments by securing and restricting access to underlying data. Strict controls and carefully designed interfaces present a managed view. Unfortunately, it is not so simple to protect private data once it has been copied to test and development environments, where access controls are more relaxed. At the

“HR and IT organizations should avoid using live personal data for test purposes. It can compromise employee’s safety, privacy and confidentiality . . .”

same time, developers and testers have unique requirements for interacting with application data. Specifically, they require access to valid data to accurately test and deploy their SAP applications.

In a recent research note, Gartner cautions against using personal data for testing purposes, particularly for HR and payroll systems that contain the data items needed for identity fraud. “HR and IT organizations should avoid using live personal data for test purposes. It can compromise employees’ safety, privacy and confidentiality, and such use is considered illegal under European Union (EU) data protection regulations.”²

You should now be asking, “Do SAP non-production environments really need to contain production data?” The answer is “No.” Gartner and other industry analysts also concur that as a best practice, masking or de-identifying the data is a viable approach. De-identifying data in non-production environments is simply the process of systematically removing, masking or transforming data elements that could be used to identify an individual. “Using dummy or scrambled data, organizations can avoid the risk of compromising personal data.”³

Selecting a comprehensive data privacy solution

Protecting data privacy is no longer optional — it’s the law! SAP customer sites must have procedures in place to manage this data across non-production environments and still comply with data privacy regulations. Effective privacy protection strategies ensure the confidentiality of private information and improve the security across your non-production environments. But, what capabilities should you look for in an enterprise data privacy solution?

As a recognized best practice, de-identifying data provides the most effective way to protect privacy and support compliance initiatives. The capabilities for de-identifying confidential data must allow you to protect privacy, while still providing the necessary “realistic” data for use in development, testing, training or other legitimate business purposes. Look for a data privacy solution that provides:

- **Comprehensive data masking techniques.** While scrambling techniques mask some data, other data, such as bank codes and account numbers, must be fictionalized and contextually valid. The ideal data privacy solution must provide a variety of masking techniques. Some of the simplest techniques may mask character or numeric data or generate random or sequential numbers, while more advanced masking routines can be used to support complex data privacy requirements.
- **Support for SAP application logic.** Data masking techniques must respect the application logic and make sense to the person viewing the results, that is, the masked data should resemble the original information. Numeric fields should retain the appropriate structure and pattern and must remain within a range of permissible values, so that functional tests pass all application validity checks.
- **Support for business context data elements.** Data masking techniques must include capabilities that respect the business context of specific data elements. For example, prepackaged capabilities for accurately masking Social Security numbers, credit card numbers and e-mail addresses would be a definite advantage.
- **Capabilities that preserve the data integrity.** Data masking techniques must preserve the referential integrity of the data. Look for capabilities that automatically mask and propagate masked data elements accurately across related infotypes, as well as applications, databases, operating systems and hardware platforms to support valid results. If the solution does not preserve the integrity of the data, processing results will be inaccurate.

In short, you need a data privacy solution that can scale to meet your current and future enterprise data masking requirements.

Meeting data privacy challenges with IBM Optim

The IBM Optim Data Privacy Solution for SAP Applications provides comprehensive capabilities for de-identifying application data that can be used effectively across non-production environments. Optim's data masking technology preserves the integrity of the data and produces consistent and accurate results that reflect the application logic.

Masked data can be propagated accurately across multiple non-production environments to generate valid results. Lastly, Optim's data masking techniques are scalable and can be deployed across applications, databases, operating systems and hardware platforms to meet your current and future needs. Optim enables organizations to meet even the most complex data privacy challenges by providing the fundamental components of effective data masking.

Application-aware data masking. Optim's application-aware data masking capabilities understand, capture and process SAP application data elements accurately so that the masked data does not violate application logic. For example, surnames are replaced with fictionalized surnames, not with meaningless text strings. Numeric fields retain the appropriate structure and pattern. Similarly, if employee ID numbers are four digits, and range in value from 0001 to 1000, then a masked value of 2000 would be invalid in the context of the application test. Checksums remain valid, so that functional tests pass all application validity checks.

Context-aware data masking. Optim's context-aware, prepackaged data masking routines de-identify key data elements across SAP applications, including HCM. Optim provides a variety of proven data masking techniques that can be used to de-identify many types of sensitive information, such as birth dates, bank account numbers, national identifiers (like Canada's Social Insurance numbers or Italy's Codice Fiscale), benefits information, health insurance identification numbers and so on.

Optim's Transformation Library™ routines allow for accurately masking complex data elements, such as Social Security numbers and e-mail addresses. Built-in lookup tables support masking names and addresses. You can also incorporate site-specific data transformation routines that integrate processing logic from multiple related applications and databases and provide greater flexibility and creativity in supporting even the most complex data masking requirements.

Persistent data masking. Optim's persistent masking capabilities generate transformed replacement values for source columns and propagate the replacement values consistently and accurately across applications, databases, operating systems and platforms. Persistent data masking capabilities ensure scalability for protecting privacy across multiple SAP application development, testing and training environments.

Proven data masking techniques

Optim provides a comprehensive set of proven data masking techniques to transform or de-identify data. The method you use will depend on the type of data you are masking and the result you want to achieve. Some of the masking techniques available with Optim are explained in the following paragraphs.

Masking character and numeric data. Optim provides several techniques for masking character and numeric data. At a simple level, a String Literal can be used to specify a value for masking alphanumeric data. You can define a String Literal using any combination of characters or numbers enclosed in quotation marks. For example, in an auto insurance context, it would be easy to substitute “Code60” for a settlement value of a claim. Similarly, the Substring masking technique returns a substring or portion of the content of a column. Using a substring that includes the area code and first three digits of the phone number provides the needed details and prevents access to actual phone numbers.

The Sequential masking technique can be used with character or numeric data types and returns a value that is incremented sequentially. For example, this technique can be used to mask checking account numbers in a banking application by simply specifying a starting account number and then incrementing each number by seven.

The Random masking technique returns a value selected at random from within a range of user-specified values and can be used to mask character or numeric data. For example, in testing a health insurance application, random numbers can be generated to mask the subscriber ID, the group number, the card number, card date and payer ID.

The Shuffle masking technique provides the ultimate in random data masking. This technique redistributes data from a single or multiple columns among a specified number of rows and optionally enforces uniqueness across shuffles. You can apply the technique to virtually any type of data, and it can be easily used to mask first names, last names, or both first and last names, and address information, including street address, city, county and postal or ZIP code.

Masking data using lookup values. Another approach to de-identification is to transform data using lookup substitution values. You can use the Lookup technique to mask a value in a source column by returning a corresponding masked value to a destination column. For example, a lookup table might transform medical diagnostic codes into fictionalized codes for testing purposes.

The Random Lookup technique allows for masking a value from a source column by returning a corresponding masked value selected at random to a destination column. Optim provides several predefined lookup tables that increase the ease of masking data:

- **First names lookup.** Contains more than 5,000 first names for de-identifying personal information.
- **Last names lookup.** Contains more than 80,000 last names for de-identifying personal information.
- **Street address/city/state/ZIP code lookup.** Contains more than 100,000 US locations to mask complete address information.

An enhanced Random Lookup technique makes it easy to transform data in any or all columns of a row in a destination table by replacing it with an entire row of data randomly selected from a lookup table. For example, instead of substituting one ZIP code for another, this feature makes it possible to mask an entire street address, city, state and postal or ZIP code.

Masking sensitive data using Optim's Transformation Library. Optim's Transformation Library makes it possible to generate valid, masked values to de-identify Social Security numbers, credit card numbers and e-mail addresses:

- **Social Security numbers.** Generates valid, transformed numbers that follow the rules used by the US Social Security Administration. For example, this feature can be used to mask Social Security numbers in testing an application that processes unemployment benefits.
- **E-mail addresses.** Generates valid, transformed e-mail addresses using string literals or the first/last name columns and the domain. For example, this feature can be used to mask e-mail addresses in a direct marketing application used to train new employees.

Preserving the integrity of the masked data. Each of the methods described so far is effective for masking data to safeguard confidentiality. However, with relational database applications, such as SAP HCM, there is an added complication. Specifically, you need the capability to propagate a masked data element to all related infotypes in the database to maintain the referential integrity of the data.

Optim provides full support for key propagation, allowing you to assign a value to a primary key or foreign key column and propagate that value to all related tables. The value you specify can be a valid column name, string literal, expression or other masked value. For example, consider two related SAP HCM tables (Figure 1). The person infotype PA0002 is parent to the address infotype PA0006, and its primary key column, *PersID*, is a 5-digit numeric value. The *PersID* represents a foreign key in PA0006.

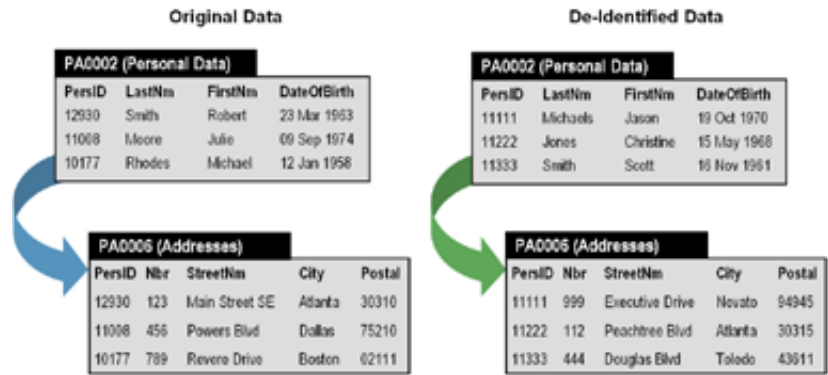


Figure 1. Optim's Key Propagation capability helps preserve referential integrity, even when data is masked.

In the Figure 1 example, the *PersID*, *LastNm*, *FirstNm* and *Date Of Birth* columns in the personal infotype PA0002 are all masked. While in the address infotype, PA0006, the *PersID* and *City* columns are masked. Note that the masked values for the primary key column in the personal infotype (*PersID*) are propagated to the foreign key column (*PersID*) in the Address infotype. In this way, the key relationship between the Personal infotype and address infotype in the test database remains intact. Without the capability to propagate masked values, the referential integrity of the data would be severed, thus creating orphan rows for the address infotype (Figure 2).

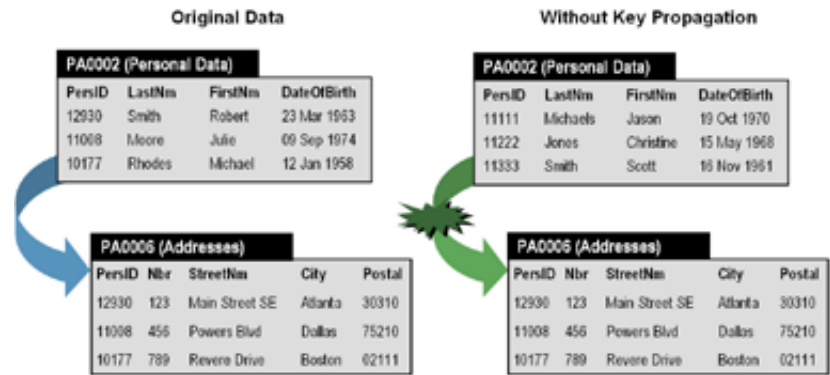


Figure 2. Without a Key Propagation capability, critical data relationships would be severed.

The capability to accurately propagate key values helps preserve the referential integrity of the test database to support valid test results. Imagine the complexity when there are hundreds of related tables involved, and keys must be propagated to all related infotypes. Without a propagate capability, many orphan tables would result, and the test database would easily become corrupted.

User-defined masking routines. When you need to perform more complex data transformations, you can prepare user-defined exit routines. These are simply programs or sets of instructions that perform the desired data transformation. Exit routines are especially useful for generating values for destination columns that cannot be defined using any other method. For example, it may be necessary to generate a value for the Customer ID code, based on the customer’s geographic location, average account balance and volume of transaction activity. The Customer ID code generated using this exit routine is then used to populate a destination column.

Data privacy best practices summary

The need to protect the privacy and confidentiality of sensitive data spans production and non-production application environments, across industries and geographic boundaries. And while many companies have implemented effective measures to protect data in SAP production environments, they are just beginning to turn their attention to the vulnerabilities in the non-production environments.

However, there are many challenges because the protective measures that apply in production environments do not necessarily support the needs of non-production environments. Development, testing/quality assurance and training teams need realistic data to accurately perform their respective activities. De-identifying data provides a means to systematically remove, mask or transform data elements that could be used to identify an individual. Data that has been de-identified is valid and useable in non-production database environments.

The IBM Optim Data Privacy Solution for SAP Applications provides a variety of data transformation techniques and built-in lookup tables for masking context-sensitive data elements, and even supports custom data masking routines. The Transformation Library provides the capability to generate and propagate valid, masked values for Social Security numbers, credit card numbers and e-mail addresses to protect privacy, while ensuring accuracy. Most importantly, you can propagate masked data elements accurately across related infotypes to help preserve the referential integrity of the database. On a higher level, masked data can be propagated accurately across applications, databases, operating systems and hardware platforms to protect your entire enterprise.

Optim supports the leading database management systems and provides federated access capabilities that allow you to extract and mask appropriate data from various production data sources in a single process. Optim also provides a single, scalable data privacy solution with flexible capabilities that can be easily adapted to your current and future requirements. Implementing Optim helps you comply with data privacy regulations and protect the confidentiality of your sensitive information across your enterprise.

About IBM Optim

IBM® Optim™ enterprise data management solutions focus on critical business issues, such as data growth management, data privacy compliance, test data management, e-discovery, application upgrades, migrations and retirements. Optim aligns application data management with business objectives to help optimize performance, mitigate risk and control costs, while delivering capabilities that scale across enterprise applications, databases and platforms. Today, Optim helps companies across industries worldwide capitalize on the business value of their enterprise applications and databases, with the power to manage enterprise application data through every stage of its lifecycle.

For more information

To learn more about IBM Optim enterprise data management solutions, contact your IBM sales representative or visit: www.optimsolution.com.



© Copyright IBM Corporation 2008

IBM Software Group
111 Campus Drive
Princeton, NJ 08540-6400
U.S.A.
www.optimsolution.com

Produced in the United States of America
10-08
All Rights Reserved.

¹ Richard Mogul, "Danger Within – Protecting your Company from Internal Security Attacks," *Gartner*, August 2002.

² Thomas Otter, "Testing Times for HR Systems and EU Data Protection Law," *Gartner Research Publication*, ID Number: G00157833, 6 June 2008, p.1.

³ *Ibid.*, p. 3.

IBM, the IBM logo, Optim and the Transformation Library are trademarks or registered trademarks of the IBM Corporation in the United States, other countries or both. All other company or product names are trademarks or registered trademarks of their respective owners.

References in this publication to IBM products, programs or services do not imply that IBM intends to make them available in all countries in which IBM operates or does business.

Each IBM customer is responsible for ensuring its own compliance with legal requirements. It is the customer's sole responsibility to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer is in compliance with any law.