Integrated Data Management Solutions

March 2009

Information Management

Closing the data privacy gap: Protecting sensitive data in non-production environments

Contents

Executive summary

This white paper explains why protecting sensitive 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 response to this issue, 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 ERP, CRM and custom 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 sensitive data, including employee and customer information, as well as corporate confidential data and intellectual property? "De-identifying" or masking data is recommended 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 the simplest and most complex privacy requirements. These techniques must produce results that accurately mask context-sensitive data in a way that reflects the application logic and preserves the integrity of the data. The solution must also include a way to propagate masked data elements consistently across applications and operating environments to generate valid results.

The IBM® Optim™ Data Privacy Solution provides a comprehensive set of data masking techniques that can support your data privacy compliance requirements:

- *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 payment card numbers, Social Security numbers 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. Critical business applications typically collect this information for legitimate purposes; however, given the interconnected nature of the Internet and information systems, as well as enterprise ERP, CRM and custom business applications, sensitive data is easily 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 receives global attention. Data privacy is a global issue that crosses applications, databases, operating systems and hardware platforms. The following table includes examples of international and industry-specific standards and data privacy regulations.

Industry	Privacy standard
Banking/Retail	Payment Card Industry Data Security Standard (PCI DSS)
Country	**Privacy legislation**
Australia	Privacy Amendment Act of 2000
Canada	Personal Information Protection and Electronic Documents Act
European Union	Personal Data Protection Directive of 1998
New Zealand	Privacy Act of 1993
Hong Kong	Hong Kong Personal Data (Privacy) Ordinance of 1995
United Kingdom	Data Protection Act of 1998
United States	Gramm-Leach-Bliley Act of 1999 Health Insurance Portability and Accountability Act of 1996

Although the specifics may differ, these data privacy regulations have a number of common elements. First, they are designed to protect individuals against the misappropriation and misuse of personal information. Second, the data privacy laws are complex, making compliance difficult and often requiring changes to corporate policies and operating procedures, as well as the adoption of new technologies.

Finally, although enforcement may focus on education and remediation, the laws generally impose substantial penalties for non-compliance—especially in cases of criminal misconduct. Interestingly enough, many laws do not specify technology requirements for compliance. Yet failure to comply with these regulations can have a significant impact.

Data privacy compliance affects your business. Protecting data privacy is a critical business initiative. A high percentage of data breaches are actually from internal weaknesses. Examples range from employees, who may misuse payment card numbers and other sensitive information, to those who save confidential data on laptops that are subsequently stolen. Lastly, outsourcing application data to offshore processing environments makes it difficult to control access to otherwise unsecured sensitive data and comply with Safe Harbor directives.

The stakes are high. Corporations and their officers face fines that can reach US$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 sensitive information risk not only losing customers and revenue, but also risk going out of business. A Ponemon Institute study of retail banking customers suggests that consumers will only do business with retailers that can protect their privacy. The study indicates that it would take only a single privacy breach for 34 percent of customers to cease doing business with a company. That number goes up to 45 percent when personal information is breached twice. According to the study, "Trust translates into strong brand loyalty, but the lack of confidence in a bank's data security capability may result in significant customer attrition."[1]

If losing customers does not have a financial impact on your business, consider that investigating a breach can cost millions. Another Ponemon Institute benchmark study, *The 2008 Annual Study: Cost of a Data Breach* for the United States, offers some staggering results: "Breaches included in the survey ranged from less than 4,200 to 113,000 records from 17 different industry sectors."[2] The study's key findings indicate that "the total average costs of a data breach grew to $202 per record compromised, an increase of 2.5 percent since 2007 ($197 per record) and 11 percent compared to 2006 ($182 per record). Breaches are costly events for an organization; the average total cost per reporting company was more than $6.6 million per breach (up from $6.3 million in 2007 and $4.7 million in 2006) and ranged from $613,000 to $32 million."[3]

Closing the data privacy gap in non-production environments

Depending on the industry, operations and types of applications, many production and non-production databases will process sensitive 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.

Corporations must protect all sensitive information, whether it resides in a production system or non-production (development, testing and training) environment. Most production environments have established security and access restrictions to protect against data breaches. Standard security measures can be applied at the network level, the application level and the database level. Physical entry access controls can be extended by implementing multifactor authentication schemes, such as key tokens or even biometrics. However, these protective measures cannot be simply replicated across every environment because the methods that protect data in production may not necessarily meet the unique requirements for protecting non-production environments.

What makes non-production environments so vulnerable? The answer lies in the nature of how non-production databases are created and used. For example, applications must be tested outside the production environment so that when testing reveals application errors, the live production system is not affected. Realistic data is essential for testing application functionality and to ensure accuracy and reliability. However, using real data increases the risk of that data falling into the wrong hands and being susceptible to a breach.

For example, while functional testing an online banking application, a tester can log in as a system user, process transactions and view customer account numbers, Social Security numbers and other confidential information. Even though this "real" data resides in the non-production testing environment, it is now at risk for being misused or misappropriated.

Similarly, application stress/load/performance tests are designed to simulate hundreds or thousands of users accessing an application database simultaneously. In this scenario, real data is often extracted from a production environment and loaded into an automated testing tool like IBM Rational® Performance Tester, then saved in a data pool to automate the instantiation of the test. This data could potentially remain accessible in an unprotected, non-production environment. If that data is captured and subsequently falls into the wrong hands, it could mean disaster for an organization.

It is recommended that production data and databases be encrypted. However, even if the data is encrypted in the production environment, it is not completely safe. Once that data is exported from the production database (into a spreadsheet or some other file format), the encryption is no longer valid, and the data is at risk of being lost or stolen.

In addition, adequate application development, testing or training coverage can require multiple environments. Most often, one or more non-production environments can be created by simply cloning copies of the production database. By definition, this means that sensitive information is propagated from a secure production environment to one or more vulnerable non-production environments.

How can organizations reduce risk? You should now be asking, "Do non-production environments really need to contain production data?" The answer is "No." Industry analysts recognize that data privacy in application development, testing and training, as well as other non-production environments, is essential. They 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. Data de-identification enables developers, testers and trainers to use realistic data and produce valid results, while still complying with privacy protection rules. Data that has been scrubbed or cleansed in such a manner is generally considered acceptable to use in non-production environments and ensures that even if the data is stolen, exposed or lost, it will be of no use to anyone.

Selecting a comprehensive data privacy solution

Protecting data privacy is no longer optional—it's the law! Companies must have procedures in place to protect data across both production and non-production environments to comply with data privacy regulations and avoid risk. Effective privacy protection strategies ensure the confidentiality of private information and improve security across your database 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 of and support compliance initiatives in non-production environments. The capabilities for de-identifying confidential data must allow you to protect privacy while still providing the necessary "realistic" data for development, testing, training or other legitimate business purposes. Look for a data privacy solution that provides:

- ***Comprehensive data masking techniques.*** *The ideal data privacy solution must provide a variety of easy-to-use masking techniques. Some of the simplest techniques may mask character or numeric data, or generate random or sequential numbers. More advanced masking routines can be used to support complex data privacy requirements.*
- ***Support for 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, payment 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 tables, as well as applications, databases, operating systems and hardware platforms, to ensure valid test results. If the solution does not preserve the integrity of the data, testing results will be inaccurate.*

- ***Flexibility, scalability and adaptability.*** *Data masking capabilities must enable extracting data from a production environment and masking the data before it is inserted or loaded into a non-production destination database. This capability ensures that real data is never propagated outside the secure production environment. In addition, in cases where cloned production databases have already been created, you need capabilities to mask data "in place." This capability ensures that you can protect sensitive data no matter where it resides. Lastly, data masking capabilities must be scalable across applications, databases, operating systems and hardware platforms to adapt to your changing requirements.*

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 provides comprehensive capabilities for de-identifying application data that can be used effectively across non-production environments. Optim 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 test results. For example, your production database can reside on IBM DB2® for z/OS®, while your testing environment resides on an Oracle® database running Linux®. Lastly, Optim data masking techniques are scalable and can be deployed across applications, databases, operating systems and hardware platforms to meet your current and future needs.

Fundamentals of effective data masking

Optim enables organizations to meet even the most complex data privacy challenges by providing the following fundamental components of effective data masking.

Application-aware data masking. Optim's application-aware data masking capabilities understand, capture and process data elements accurately so that the masked data does not violate application logic. For example, surnames are replaced with random surnames, not with meaningless text strings. Numeric fields retain the appropriate structure and pattern. For example, if diagnostic codes 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. Most importantly, Optim propagates all masked data elements consistently throughout a test database and to other related applications and databases.

Context-aware data masking. Optim's context-aware, prepackaged data masking routines help de-identify key data elements. 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 Transformation Library™ routines allow for accurately masking complex data elements, such as Social Security numbers, payment card 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 capability generates transformed replacement values for source columns and propagates the replacement values consistently and accurately across applications, databases, operating systems and hardware platforms. Persistent data masking capabilities ensure scalability for protecting privacy across multiple development, testing and training environments.

Proven data masking techniques

Optim provides a comprehensive set of 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. For example, it is easy to mask customer identification numbers by simply applying a random or sequential number masking technique, or by replacing patient names with a predefined string of text. 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 company, it would be easy to substitute "Code60" for a text description of the 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 a character or numeric data type and returns 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 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, as well as address information including street address, city, county and postal or ZIP code.

In addition, you can easily mask date-driven information, such as birth and death dates or admission and discharge dates, device identifiers and/or serial numbers and Web addresses or URLs and Internet Protocol (IP) addresses. Adding to these types of data, you can use the Shuffle technique to mask any type of account numbers, medical record numbers, health plan and beneficiary identifiers, certificate or license numbers, vehicle identifiers, employee identification numbers and payment card numbers.

Masking combined values. Data masking can also be achieved using concatenated expressions. These expressions enable you to mask the value of a destination column by combining the values of two or more source columns or by combining a column value with some other value.

For example, suppose that bank account numbers are formatted "999-9999," where the first three digits represent the type of account (checking, savings, money market and so on) and the last four digits represent the customer identifier. Here, you could mask the account number by concatenating a substring using the first three digits of the actual account number with a four-digit number derived using the sequential masking technique. In this example, concatenation allows you to retain the correct format of the account number column, preserve important information about the type of account and, at the same time, de-identify the confidential customer information. The result is a fictionalized account number that is still valid in the context of the application test.

Masking data using lookup values. Another approach to de-identification is to transform data using 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 masking technique allows for masking a value from a source column by returning a corresponding masked value selected at random for 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 U.S. 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/ZIP code.

Masking sensitive data using Optim's Transformation Library. Optim's data 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 United States Social Security Administration. For example, this feature can be used to mask Social Security numbers in testing an application that processes unemployment benefits.*
- ***Credit card numbers:*** *Generates valid, transformed numbers, based on rules used by credit card issuers. For example, to support PCI DSS compliance, this feature can be used to mask payment card numbers in testing an application that processes customer statements.*
- ***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, there is an added complication. Specifically, you need the capability to propagate a masked data element to all related tables in the database in order to maintain referential integrity.

If a masked data element (for example, a telephone number) is a primary or secondary (foreign) key in a database table relationship, then this newly masked data value must be propagated to all related tables in the database. Key propagation helps preserve the referential integrity of the transformed data. Without key propagation, the relationships between parent and child tables would be severed, causing the test data to be inaccurate. Consequently, application testing will produce unreliable results.

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 masking technique. Assume a simple example consisting of two related tables: Customers and Orders (see Figure 1). The Customers table is parent to the Orders table and its primary key column, *Cust_ID*, is a 5-digit numeric value.

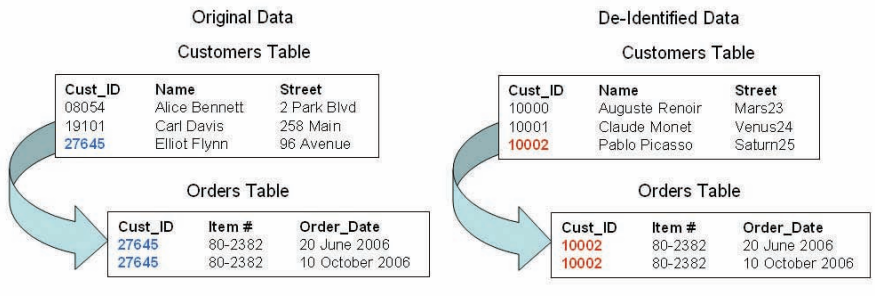
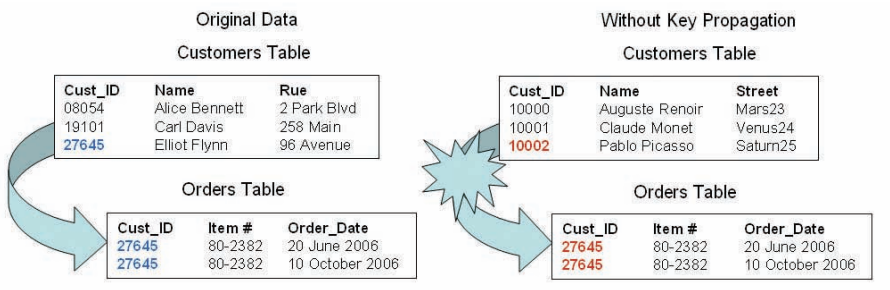


Figure 1: Optim's key propagation capability ensures referential integrity, even when data is masked.

In the Figure 1 example, the *Cust_ID*, Name and Street columns are masked. Notice that the name "Elliot Flynn" has been masked as "Pablo Picasso." The street name has also been masked. In particular, the sequential masking technique was used to transform the original *Cust_ID* for Elliot Flynn from 27645 to 10002. When this masked *Cust_ID* value is propagated from the Customers table (parent) to all related tables, the key relationship between the Customers and Orders tables in the test database remains intact. Without the capability to propagate masked values, the referential integrity of the data would be severed, creating orphan rows for the Orders table (see Figure 2).

Figure 2: Without a key propagation capability, critical data relationships would be severed.



The capability to propagate key values helps preserve the referentially 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 tables. 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, a tester may need to round year-to-date sales revenue before supplying test data to an application reporting process. In another scenario, 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.

Mask and Move or Mask in Place. With Optim, you can prevent production data from every being exposed outside a secure production environment. Optim's "Mask and Move" capabilities allow you to extract and mask data and then insert or load the data into one or more destination non-production databases.

In addition, Optim's "Mask in Place" capabilities allow you to de-identify data extracted using third-party tools and de-identify data that already resides in cloned non-production environments. These options provide flexibility for organizations that have data in place for testing, or that use backup facilities to create those test databases. Using Optim to mask data directly where it resides eliminates the need to move data for additional processing and still preserves the referential integrity of the data.

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 production environments, they are just beginning to turn their attention to the vulnerabilities in the non-production database 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 support 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 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, payment card numbers and e mail addresses to protect privacy while ensuring testing accuracy. Most important, you can propagate masked data elements accurately across related tables 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 Integrated Data Management Solutions

IBM Optim Integrated Data Management Solutions offer proven, integrated capabilities to manage enterprise application data from requirements to retirement. With Optim, teams can share data artifacts (like models, policies and metadata) to align data management with business goals and improve collaboration. Today, organizations of all types leverage Optim to improve performance, streamline database administration, speed application development, and enable effective governance. Optim delivers better business outcomes, at lower cost, with less risk, while providing capabilities that scale across enterprise applications, databases and platforms.

For more information

To learn more about IBM Optim Integrated Data Management Solutions, contact your IBM sales representative or visit **ibm.com**/software/data/data-management/optim/data-privacy-solution

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

www.optimsolution.com

Produced in the United States of America
March 2009

[1] "2006 Privacy Trust Study for Retail Banking," The Ponemon Institute, LLC and Vontu, Inc., January 2006, as referenced in "Ponemon Institute Names Most Trusted Retail Banks," Vontu Press Release, January 26, 2006.

[2] "2008 Annual Study: Cost of a Data Breach," United States, PGP Corporation, 2009. Benchmark research conducted by the Ponemon Institute, LLC. Approved for redistribution by The Ponemon Institute, p. 4

[3] Ibid., p. 4

IMW14008-USEN-01