Data masking strengthens privacy and security by changing sensitive data while maintaining its integrity.

Data Masking: Strengthening Data Privacy and Security

So, you are responsible for making sure that sensitive data is seen only by those who need it. You've got a lot of bases covered – laptops are encrypted, network access controls are established, and database activity is being monitored. But you know there are still a few holes – software development still gets copies of production data for testing purposes, new staff still use customer records for training, and the outsourcing company that handles customer service shouldn't be using Social Security Numbers to verify identities.

But you can plug these holes by masking your data. Data masking strengthens privacy and security by changing data while preserving its integrity. By using the original data as a model, masked data looks and behaves like real data, even though it's NOT real data. Sure, some people will still need access to the real data – like the team working on the targeted marketing campaign – but by masking anything they don't need, you limit data exposure to the amount necessary for people to do their jobs. And isn't that your ultimate goal?

Many Fortune 500 companies have already integrated data masking into their privacy and security strategy, and you can too. We'll go deeper into how your sensitive data is being exposed, why that concerns you and your organization, and how data masking can help. Read on …

Your data is exposed (but you may not know it)

"62% [of respondents] use live data for testing of applications and 62% of respondents use live data for software development."[1]

Ponemon Institute, December 2007

We can all agree that we value our privacy, and that of others, much higher than we did in the past. Yet, we live in a new era in which the proliferation of data demands that we take action on these values. Not only is there more data about us, but this data is being replicated and shared with zeal. Recently, IDC estimated that 281 billion gigabytes of data were produced in 2007, and that "enterprises are responsible for the security, privacy, reliability, and compliance of 85% [of this data]."

[1] Ponemon Institute LLC, The Insecurity of Test Data: The Unseen Crisis, December 2007, p. 3.

Moreover, IDC predicts that 1.8 trillion gigabytes of data will be created in 2011.[2]

Development processes rarely need real data – they just need something that behaves like real data.

Copies of production data are used in a lot of business processes, but one that is very common is in application development, as illustrated by the above Ponemon Institute quote. Typically, development groups will attempt to generate their own data, only to find that it doesn't adequately emulate real data. Inevitably, they request real data and a copy is made. Ironically, development processes rarely need real data - they just need something that behaves like real data.

Unlike development, many business processes do need real data. Marketing teams want to know purchase information to target campaigns. Outsourcers need to have customer information to handle inquiries.   Financial analysts need to know account information to mitigate risks. There will always be business processes that need real data to make decisions.

There will always be business processes that require real data, but they don't always need all the information they are given.

However, these processes rarely need all the information they are given. Do marketing teams need to know exactly who bought what, or do they categorize according to demographics? Do outsourcers need to have access to complete SSNs to assist in verifying identities, or do they just need the last four digits? Do analysts need to know the full account number of defaulted loans in order to determine which branches are making poor credit decisions? These examples of data exposure aren't that special – in fact, you probably have access to sensitive data which you don't need in order to perform your job.

Why should my company care about data exposure?

"The total average cost of a data breach grew to $197 per record [in 2007]."[3]

Ponemon Institute, November 2007

Risk

Risk – "possibility of loss or injury"

Merriam-Webster's Collegiate Dictionary

The financial costs associated with exposing sensitive data are absolutely staggering.   Using the above number provided by the Ponemon Institute, a data breach involving 100,000 records

---

[2] John F. Gantz et al., The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011, IDC, March 2008, pp. 2-3.
[3] Ponemon Institute LLC, 2007 Annual Study: U.S. Cost of a Data Breach, November 2007, p. 2.

would cost $19.7 million. It doesn't matter how big your organization is, $19.7 million is a lot of money.

In case you think the Ponemon number is too high to believe, you may want to know that Forrester Research estimates the cost of <u>low-profile</u> breach in a regulated industry to be $155 per record.[4] Using the Forrester numbers, such a breach involving 100,000 records would cost $15.5 million.

The cost of lost business due to a data breach is $128 per record.
- Ponemon Institute

Data breaches incur costs from a variety of sources including legal, forensics, auditing, consulting, brand damage, public relations, customer mailings, and free credit monitoring services. Yet, the most shocking cost is due to lost business, which the Ponemon Institute has determined to be $128 per record.[5] It should also be noted that U.S. courts are mixed on the idea of financially awarding victims of privacy violations, so the risk of such restitution may increase in the future.

Compliance

Compliance – "conformity in fulfilling official requirements"

Merriam-Webster's Collegiate Dictionary

If the risks haven't convinced you, then you should consider the compliance angle. We present a few of the relevant U.S. regulations and standards below.

GLB

The Gramm-Leach-Bliley Act offers Privacy and Safeguards Rules to protect personal information held by U.S. financial institutions. The Privacy Rule speaks largely to information collection and sharing – with respect to data exposure, this rule mandates that certain information, such as account numbers, cannot be shared with third parties. The Safeguards Rule speaks more to protecting information.

GLB Act oversight is conducted by the U.S. Federal Trade Commission, who has begun to pursue violators more aggressively. Recent settlements regarding data exposure have online advertiser ValueClick paying $2.9 million[6], and retailer TJX subject to 20 years of FTC scrutiny.[7]

---

[4] Khalid Kark with Paul Stamp, Jonathan Penn, and Allissa Dill, <u>Calculating The Cost Of A Security Breach</u>, Forrester Research, Inc., 10 April 2007, p. 4.
[5] Ponemon Inst., <u>Breach</u>, p. 2.
[6] U.S. Federal Trade Commission, "ValueClick to Pay $2.9 million to Settle FTC Charges" (news release), 17 March 2008.
[7] U.S. Federal Trade Commission, "Agency Announces Settlement of Separate Actions Against Retailer TJX, and Data Brokers Reed Elsevier and Seisint for Failing to Provide Adequate Security for Consumers' Data" (news release), 27 March 2008.

HIPAA

The Health Insurance Portability and Accountability Act offers Privacy and Security Rules to protect personal health information held by U.S. healthcare organizations.   The Privacy Rule makes explicit references to limiting information to the minimum required to accomplish a purpose.

"Good privacy is good business"
- Karen Curtis, Privacy Commissioner of Australia

HIPAA is enforced by the U.S. Department of Health and Human Services. Criminal penalties can be as high as $250,000 and include prison sentences of up to 10 years.[8] Recently, Providence Health & Services agreed to pay $100,000 and implement a "Corrective Action Plan" to resolve HIPAA data exposure violations.[9]

PCI DSS

The Payment Card Industry Data Security Standard is a set of requirements for securing payment account data. The PCI DSS affects all the companies which handle payment card data, which are myriad. The requirements are straightforward, and include "protect stored cardholder data" and "restrict access to cardholder data by business need-to-know".[10]

Recently, the credit card processor associated with the TJX data breach was fined $880,000 for failing to meet this standard.[11] In the same incident, TJX paid a $40.9 million settlement to Visa.[12]

Why should I care about data exposure?

"Data Breach Fallout: Do CISOs Need Legal Protection?"

Recent headline from CSO Security and Risk

Accountability

Accountability – "an obligation or willingness to accept responsibility"

Merriam-Webster's Collegiate Dictionary

Generally, data breaches compromise the job security of those who are accountable. In 2006, AOL fired a researcher and their

---

[8] Office for Civil Rights, U.S. Department of Health and Human Services, <u>Summary of the HIPAA Privacy Rule</u>, May 2003, p. 18.
[9] Office for Civil Rights, U.S. Department of Health and Human Services, "HHS, Providence Health & Services Agree on Corrective Action Plan to Protect Health Information" (news release), 15 July 2008.
[10] Payment Card Industry Security Standards Council, <u>Payment Card Industry Data Security Standard</u>, Version 1.1, September 2006.
[11] SC Magazine, "Visa fines TJX credit card processor", www.scmagazineus.com, 29 October 29 2007.
[12] SC Magazine, "TJX agrees to $41 million settlement with Visa", www.scmagazineus.com, 30 November 2007.

supervisor for exposing search data on the web; in the same incident, the CTO resigned.[13] In the best case scenario, those who are accountable must apologize for the errors of others, as was the case when U.S. Secretary of State Condoleezza Rice apologized for incidents of snooping into the passport files of Senators Obama, McCain, and Clinton.[14]

The U.K. and Hong Kong are considering criminalizing data abuses.

Recently, there has been discussion of the need for data breach liability protection for corporate officers, as demonstrated by the above CSO Security and Risk headline. Moreover, the U.K. and Hong Kong are both considering criminalizing data abuses.[15][16]

Data masking limits exposure of sensitive data

"We're not going to solve this by making data hard to steal.
We're going to solve it by making the data hard to use."

Bruce Schneier, security expert

Data masking ensures that the end-user is given only the information that they need to see.

Data masking is referred to by several different terms, including data obfuscation, data de-identification, and anonymization. Regardless of the term you use, the concept is simple – data masking changes your data so that only the necessary information is exposed to the end-user.

Data masking is not intended to replace existing privacy and security technologies – it is intended to augment them. Each technology serves a specific purpose: authentication ensures that only the right people access the database, antivirus software ensures that rogue programs aren't transmitting the data, and database activity monitoring ensures that sensitive data isn't being accessed unnecessarily. Data masking ensures that the end-user is given only what they need to see.

Data Masking and Encryption

"We continue to urge individuals and organizations to take basic data security precautions such as … technologies which enhance security and privacy such as data encryption and anonymizing (data masking) services."[17]

Privacy Commissioner of Canada, 2007 Report

[13] Anick Jesdanun, "3 employees leave AOL in search data fallout", Associated Press, 21 August 2006.
[14] Anne Flaherty, "Passport files for Clinton, Obama, McCain were pried into by State Department workers", Associated Press, 21 March 2008.
[15] United Kingdom, House of Commons, Protection of Private Data, Justice Committee, 3 January 2008.
[16] Hong Kong, Office of the Privacy Commissioner for Personal Data, "The Privacy Commissioner's clarification on 'criminalizing data leakage'" (press release), 22 May 2008.
[17] Jennifer Stoddard, Privacy Commissioner of Canada, Annual Report to Parliament 2007, June 2008, pp. 40-41.

Encryption

Original: Alan Smartworth,
42 Heritage Place, 09863

Transit: d09f8lkdsf90ldf098lkf

End-user: Alan Smarworth,
42 Heritage Place, 09863

Data Masking

Original: Alan Smartworth,
42 Heritage Place, 09863

Transit: Thomas Martin,
16 Lawn Street, 09732

End-user: Thomas Martin,
16 Lawn Street, 09732

So, data masking is like encryption, right? Well, no, there is a fundamental difference. When data is encrypted, it cannot be used until it is decrypted; once it is decrypted, the original data is available. Encryption protects the data while in transit, but does not protect it from abuse at the final destination. In contrast, data masking protects the data from abuse at the final destination by restricting to only what is needed.

So, is my masked data still valid? Yes, it remains valid for its intended use. Naturally, whenever you change data you reduce its authenticity, but that is the whole point. The masked data will be sufficiently realistic to ensure that the end-user can draw the same conclusion (be it test results or analysis), while exposing them to only what they need.

So, if my data is masked a lot, do I still need to encrypt it? Well, that depends. If you feel that your specific masked data product would be benign if it got into the wrong hands, then you wouldn't need to; otherwise the masked data can also be encrypted. Just as encryption exempts lost data from breach notification laws, the same argument can be made for masked data, depending on the extent that the original data was masked.

Encryption and data masking both allow the end-user to do their required work. Encryption protects privacy in the event of theft, but does not protect privacy at the intended destination. Data masking protects privacy at the final destination, and can protect privacy in the event of theft. Encryption and data masking are complementary technologies.

Where is data masking being used?

"Viacom, YouTube agree to mask user data"

CNET News headline, July 2008

In 2007, Viacom sued YouTube alleging copyright infringement. In July 2008, courts ordered YouTube to surrender records of user data, including usernames and IP addresses. Initially, YouTube refused to release data as it would compromise the privacy of its subscribers, but later agreed to release the data provided it was masked to remove personal information.

The U.S. Census Bureau masks data all the time. They are mandated to not reveal any private information in the data products they release. Some of the ways they mask data are more sophisticated than others: they might remove the last 3 digits of a zip code, remove salaries that are exceptionally high, or change numeric values while preserving a mean.

Effective data masking ensures that masked data is realistic.

Privacy conscious software development and testing groups regularly perform primitive data masking. When they need test data, they will take the real data and blank out sensitive information. Account numbers will be replaced with '11111111', names will be replaced with 'xxxxx' or 'John', Visa numbers will be overwritten with '4500 0000 0000 0000'. This type of <u>primitive</u> data masking preserves privacy, but destroys the validity of the data; <u>effective</u> data masking is more sophisticated.

What should I look for in a data masking solution?

"Selection criteria [for data masking] should include practical implementation issues: multiplatform and multidatabase solutions; the realism of masked, quasi-real data; and workflow-driven processes."[18]

Gartner, July 2008

Data masking focus

Look for a solution that is built for data masking.

First, look for a solution which is built for data masking. Solutions which began as archiving or subsetting tools, then were modified to allow masking, have way too much overhead – this overhead makes a big difference when a 2 hour mask takes 10 hours.

Validity

You should be able to mask exactly the way you want, but you shouldn't have to perform a custom mask every time. Look for a solution that makes it easy to do replacement (replacing all values with 'Jane'), shuffling (permuting the values), and loading (using external values), as well as more complicated things like credit card generation (with valid prefixes and check sums!).

Platforms and Databases

Look for a solution that doesn't require different products for different database management systems (DBMSs). Ensure that it can handle a variety of DBMSs, and don't settle for products that offer full functionality for certain DBMSs, but limited functionality for others.

Database concepts

Look for a solution that can correctly manage key database concepts, such as filters (e.g., Mastercard only), groups (e.g., credit card types), and relational integrity. And, yes, it MUST be able to mask primary keys!

---

[18] Joseph Feiman, <u>Data Obfuscation (Masking, Privacy, Scrambling): Many Names for the Same Technique</u>, research note G00157661, Gartner, Inc., 29 July 2008.

Reusable Configurations, Batching, and Deployment

Look for a solution that doesn't require you to reconfigure when you switch databases, workspaces, or schemas. As well, you should be able to run a mask as a batched/scheduled process, possibly on a server with no graphical ability.

Consistent Masking

Finally, if you want to get the most out of your solution, make sure that there is a way to mask to consistent values on different databases/systems. If you mask 'Bob Grandy' to 'Marc Smith' in your HR data, it will present a problem when 'Bob Grandy' is masked to 'Charles Lopez' in your insurance data.

Camouflage®
DATA MASKING SPECIALISTS   **The data masking specialists.**

Enterprise-wide

Camouflage is the single-product, platform-independent solution. It offers consistent functionality across more than 10 DBMSs and handles data flow seamlessly. The patent-pending Translation Matrix functionality guarantees consistent masking across different systems and over time. For these reasons, Camouflage is <u>the</u> solution for enterprise-wide data masking.

Architecture

Camouflage was built for data masking, so it lets you mask the way you want. Camouflage offers over 20 pre-packaged masking algorithms, but also supports custom scripting.

Camouflage understands databases and leverages the patent-pending Insert Engine process to mask huge volumes of data. It can handle filters, grouping, and relational integrity with ease.

Ease of Use

Camouflage is quick to deploy and easy to use: it can be up and running on the most complicated networks in less than a week. More importantly, our customers tell us that Camouflage has a much shorter learning curve, the GUI is easier to use, and has better "flow" than solutions offered by competitors.

To see how Camouflage can strengthen your data privacy and security, schedule a demo today. Visit www.datamasking.com or contact our Sales Team at +1 866.345.8888