Managing Data Privacy Risks When Using Production Data

A White Paper from Direct Computer Resources, Inc.
September 2012

Direct Computer Resources, Inc.

Direct Computer Resources, Inc.
120 Birch Road Franklin Lakes, NJ 07417
Contact: Bill Vitiello, Principal/EVP Sales & Marketing: 877.704.0077
info@DataVantage.com

In spite of all the publicity surrounding highly visible losses of customers' and clients' Personally Identifiable Information (PII), companies, educational institutions, non-profits and government agencies persist in using processes that perpetuate the risks; specifically, using copied production data for nonproduction purposes.

Here are some names you probably know: Global Payments, The California Department of Social Services and Emory University Hospital. All of them have had breaches greater than hundreds of thousands of records *in only the first quarter of 2012*. Financial, Social Services and Healthcare records were involved.

While estimates of the sources of exposed data vary, there is a general consensus that insiders given access to production data for legitimate purposes are accountable for a majority of the losses. Given the magnitude and variety of costs associated with data breaches, it is difficult to understand why companies, educational institutions, non-profits and government agencies do not move more actively to address the major causes of them.

Current Environment

Risks

The risks of managing others' personal data are ever-present and constantly changing. Just as external threats morph into new forms that require continuous review and updating of precautions, internal risks evolve as well. Malefactors' growing ability to synthesize identities from increasingly smaller and apparently insignificant pieces of disjointed information creates an increasing burden on information managers to short circuit the connections in data that might be exposed thoroughly enough to protect the identities contained therein

The fact is that there are untold thousands of people around the world who have nothing better to do than sift through stolen data looking for connections to create information of value.

Let's look at the risks in the context of a common paradigm for thinking about information management in organizations—People, Processes and Technology:

People

People-related risks include all the factors that may contribute to the erosion of information workers' resistance to compromising PII. The drive to accomplish more with less staffing, enabled by myriad new technologies and outsourced services increasingly de-emphasizes the importance of individual workers, weakening the bond between them and the organization. Working from remote locations and with less people overall may make it difficult to achieve the critical mass necessary to foster a spirit of teamwork and sense of connection that is crucial to motivating people to adhere to a value structure that impels them to take personal responsibility for the sensitive information they steward. Also, the push to save operating expenses frequently results in more experienced/expensive workers being pushed out. In addition to lowering the aggregate skill and experience levels, this creates an impediment to more junior workers 'buying-in' to the organization's values, as they become skeptical of the link between integrity and personal development and reward within the organization.

In addition to what is taking place within the internal IT staff, there is the issue of temporary, outsourced or offshore staff. Remote to the organization's major worksites, incompletely integrated into the culture and probably minimally supervised by those who are, these workers are ripe targets for outsiders looking for a way to access personal information.

Process

Process-related risks include everything about the way tasks are executed and managed that may create opportunities to compromise personal information. While Production operations are almost always tightly controlled, other IT activities may not be. Areas for concern in this regard include Systems Development, QA and Training, among others. Every one of these activities requires lifelike data to perform properly and sourcing that data often creates an exposure that can dwarf the risks associated with day-to-day operations.

> ***It has been estimated that between backup and non-production instances, an average organization can have as many as ten copies of each of its production datasets, many of them resident in environments in which production-level security precautions are not fully observed.***

Cooperative business models and third-party processing only complicate the situation. It is often difficult to know whether organizations with whom you share data are observing the same level of vigilance as you do. When a credit card payments processor loses customer data, it is the credit card provider that must deal directly with its customers to cancel and replace the compromised accounts. Parsing the nuances of who is really at fault is seldom of interest to disgruntled cardholders and may not be to regulators, either.

> ***Simply put, any organization that uses copied production data, even out-of-date data, in support of non-production activities is setting itself up to be breached. Similarly, any organization that allows its business partners to use copied production data for non-production purposes is exposing itself to the same degree, if not more.***

Technology

Technology-related risks include all of the elements of the infrastructure that contribute to the decentralized management, distribution and sharing of data, complicating the task of monitoring its use. New technologies with attractive cost/benefit ratios are coming on-line constantly, enticing organizations to employ them in order to remain competitive. Adopting new enabling technology without fully evolving internal management processes to control its use creates exposures, especially for early adopters who do not have the benefit of others' experience to inform their use of them.

Some increasingly prominent technologies that characterize this are cloud-based architectures, outsourced data centers and business intelligence data stores. Outsourced infrastructure amplifies People and Process risks by offloading responsibility for physical control of data and reduces an organization's ability to observe and monitor day-to-day operations. An increasing focus on Big Data and the proliferation of analytical data stores and warehouses creates the potential for the exposure of a critical mass of information with unusual breadth, perhaps thought to be de-identified, to someone prepared to mine it and infer identities from it.

Regulations

Regulations are being enacted at a furious rate and in overlapping jurisdictions with increasingly onerous penalties. Industry standards have been enacted in addition to governmental regulations. Some of the more well-known are:

- HIPAA, the Health Insurance Portability and Accountability Act
- HITECH, the Health Information Technology for Economic and Clinical Health Act
- PCI DSS, the Payment Card Industry Data Security Standard
- Massachusetts 201 CMR 17, the Massachusetts Data Privacy Law
- EU Data Privacy Directive, a benchmark law addressing information management in the EU that also serves as a model for similar laws in other jurisdictions

It cannot be assumed that complying with US Federal regulations will be good enough to protect an organization on a state or local level. Similarly, many foreign countries have adopted stringent regulations of their own. Simply doing internet-based business in some of these countries, even though data is processed domestically, may subject a business to them.

Most of these regulations contain some or all of the following provisions or elements:

- The geographic scope of the regulation
- What specific information is covered, what information qualifies as sensitive
- To whom the regulation applies; e.g., companies doing business with a citizen within the geographic area addressed, regardless of where the business is domiciled
- What information processing actions qualify under the regulation, which may include manuallyprocessed information in addition to electronically-processed information
- What is required of an organization that is performing qualified information processing actions
- Standards to be observed, which may include data quality and security and auditing standards, among others
- Obligations to disclose various practices before doing business with or otherwise engaging with others
- Obligations to report to or to be monitored by regulating agencies
- Limitations on the ability to transport data across geographic boundaries
- Responsibility to report data losses
- Penalties for losses and obligations to provide remedies to affected parties

Costs of Data Breaches

There are both hard and soft costs of data losses, which include:

- Fines
- Legal expenses
- Costs of remediation
- Loss of reputation and customers
- Protective services for individuals whose data was exposed

Estimates for these costs vary, but the average estimate is about $200 per record and $6.5 Million per event over the last two years.

One thing that we have observed is that there is often some flexibility in the application of various laws and the penalties and fines that are assessed. An organization that suffers a breach that has attempted to or is in the process of implementing solutions to address exposures will be more likely to be treated with some leniency. Organizations that suffer a breach and have done only the bare minimum, or less, to manage exposures to loss are likely to feel the full force of the regulators when the losses are brought to light. Further, being penetrated from the outside is generally viewed as somewhat less preventable (as long as commercial-quality precautions are in place) while lax controls that contribute to insider theft are far more difficult to rationalize.

Protective Measures—What to do and How to do it

First of all, you must identify and prioritize the assets that need protection. Data assets that are not copied and used to support non-production activities are clearly less at-risk than those that are and can be de-prioritized safely. Also, your schedule for activities that will require protected data, such as development, QA or training, should inform your priorities. Finally, designing, building and testing a protection program for a given data asset takes some time, as determined by the scope and complexity of the asset. The lead time to implement a protection process for a given asset, then, must be considered when planning for an activity that will require protected data.

What needs to be done should be determined by making a loss-adjusted risk assessment for each of your major information assets:

- What PII does it contain? What is the data's value if lost or stolen?
- How vulnerable or susceptible to loss is it?
- What would its loss cost?
- What use will be made of the protected data?

There are a variety of considerations that should determine how elaborate your protection scheme for each particular asset needs to be. For instance:

- How complex and interwoven are the logical relationships in the source data?
- What referential and logical integrity must appear in the protected data in order for it to be useful?
- Is it necessary for specific output values to appear in the protected data? This might be the case if the protected data is to be used as a regression test set in which specific entities must be represented to enable execution of automated or scripted tests.
- Must the cardinality of the protected data set be manipulated in order to protect it properly?
- Is there a need for specially-constructed replacement value sets to enable specific logical integrity in the protected data?

A robust definition of **what,** then, consists of both a macro-level portfolio view of your data assets, their protection urgency and risk profile and a micro-level view in which the specific requirements of the protection needed constitute a functional definition of the protection requirements.

How you will implement protection must be determined from among a number of options, which should be viewed through the technology solutions paradigm we used earlier: People, Process and Technology and Cost considerations:

- **People:** How much manpower and what skillsets are required to implement the solution? Do I have those people available now or will I have to go out and hire them?
- **Process:** What changes do I need to make and how much of the organization will be effected? How much will the changes impact productivity and throughput?
- **Technology:** What, if any, new tools do I need to make this happen?
- **Cost:** What are the monetary (capital and operating) and non-monetary costs of the solution?

Clearly, there is interplay among these factors that will determine the impact for your organization. Some technology solutions require casts of thousands to operate and others demand substantial changes in existing operating processes.

Finally, **What and How** have to be matched to determine a course of action. Imputing a cost to the loss of a particular data asset creates an upper limit on the solution that will be worth implementing to prevent it. Viewing all corporate data assets that contain PII as a portfolio of prospective data protection projects may ultimately suggest a spectrum of approaches that provide the ability to pair a solution to the risk and magnitude profile of the potential loss of a given class of assets.

What becomes clear from analysis of data privacy alternatives is that solution *efficiency* and *effectiveness* is everything. The easier and faster a solution is to implement and the better the protection it provides for the original data, the more applicable it will be to a wide range of your data assets, from moderately-sized, moderate risk assets to high-vulnerability, high value ones.

Data Masking—Protecting PII Against Theft by Insiders

One of the best ways to protect data that must be shared to enable non-production processes is **masking.** Masking transforms data values in such a way that the protected data conforms to all referential and logical constraints making it usable for the desired purpose but with the following two conditions: (a) values of sensitive data elements are modified and (b) relationships in the data that might allow the original data values to be inferred are broken or scrambled. If a properly masked dataset is taken, nothing of value will be lost. When a properly masked dataset is propagated within an organization, nothing is risked.

Protecting a dataset through masking is as much an art as a science. After developing a thorough understanding of the dataset's structure and content and the intended use of the protected data, the security architect can define a least-cost strategy to subset, extract and transform the data to mitigate the risks of disseminating it while maintaining its utility.

Several capabilities define a good data masking solution:

Reproducibility

- The solution must provide transformation functions that ensure that multiple instances of a given input will produce the same output. This is a basis for maintaining referential integrity among tables in a protected dataset.
- Ideally, reproducible transformations should function predictably, regardless of whether the source data comes from homogeneous or heterogeneous platforms.

Irreversibility

- Needless to say, transformation functions should result in protected values that provide no information that would facilitate reverse-engineering them back to the original values, either individually or as an entire set of transformed data.

De-identification

- De-identification is the process of replacing a source value or values with the replacement drawn from a candidate set. This is an important capability that is necessary to enable masking of structured or constrained data, such as a national ID that is a mix of numbers and letters or an email address that is composed of letters, numbers and symbols.
- Ideally, de-identification should be implemented so as to be reproducible as well as irreversible. In addition, the ability to assign de-identified values from one column to others in a record before writing it to the protected target dataset is an extremely useful feature.

Unique De-identification

- Unique De-identification is a further refinement of de-identification in which each candidate replacement value is guaranteed to be used only once. This is critical to producing a protected set in which each entity is required to have a unique ID, such as a social security number for each individual.

Support for user-written transformation functions

- No product will be able to support every possible use case. It is important that you be able to develop and implement your own transformations and incorporate them into the masking process easily. Such requirements as date aging and the ability to invoke external procedures, such as DB Stored Procedures must also be supported.
- It is also important that the solution provide the ability to integrate pre- and posttransformation processing options with built-in transformations so that source data can be transformed into a specific format, masked and transformed back into the original format prior to its being written to the target data set. For example, social security numbers stored as character strings with embedded hyphens may need to have the hyphens removed, be cast to integers, masked and then reformatted as a character string with hyphens before being written back to the target structure.

Development Accelerators

- A solution should provide facilities to speed the delivery of masking solutions:
  - The ability to read system metadata and generate solution component
  - The ability to define masking policies and apply them to the appropriate columns of a dataset to be masked. For instance, the solution should allow the user to select a transformation to be performed on Social Security Numbers, allow the solution to identify the occurrence of all columns containing them in a source data set and apply the transformation specification to them.
  - The ability to filter and perform threaded extracts on networks of source tables, in which a subset of related tuples are extracted and transformed, with no or minimal programming.

Command Line Interfaces

- A solution should integrate with enterprise scheduling and work management systems so that extraction and masking can be implemented and managed within the enterprise production framework.

Vendor support

- Although masking solutions have been around for some time, their adoption has just begun to ramp up and most organizations have little experience in this area. An experienced security architect can make all the difference in how efficient and effective your masking solution is and support from someone who has been doing it for years can make the difference between success and, well, something less.

About Direct Computer Resources, Inc.

DCR has been working in the data protection space for over ten years and has been evolving the DataVantage tools over this time to create optimally *efficient* and *effective* solutions in nearly any environment you're working in.

DCR provides tools for both mainframe and distributed environments:

Mainframe:

- DataVantage® for DB2
- DataVantage® for IMS
- DataVantage® for VSAM

Distributed, including mainframe:

- DataVantage GLOBAL®

About the Author

Howard M. Wiener is the Director of Professional Services for DCR DataVantage, Inc.

Version 8.26.2012