When Data Is a Risk

Data Loss Prevention Tools and Their Role within IT Departments

KLAUS HALLER

Klaus Haller's work focus is on IT risk, compliance testing, and test organizations. He has concrete working experience with Symantec's Data Loss Prevention tool and brings an infrastructure and operations as well as a business analysis perspective. He has been with Swisscom IT Services since 2006 and worked as a consultant with various customers, mainly in the banking industry. He is a frequent conference speaker and publishes in various magazines. klaus.haller@swisscom.com

Snowden is a reversal point for IT security and risk. Before him, many saw IT security as equivalent to a medieval town wall: keeping outside hackers and malicious code away from the company. Firewalls, virus scanners, and application security testing (e.g., to find SQL injections) fit the town wall approach. But Snowden was different. He was from the inside of the organization. He collected large amounts of sensitive data. Then, he got the data out of a highly secured IT organization, which had to learn from the press about the case. In this article, I will explain such data-related risks in IT departments and how data loss prevention (DLP) tools help to manage them.

Understanding the Business Risks

Computer professionals think in terms of technical components: operating systems, applications, and databases. In contrast, data-related risks require a business view. First, there is the risk of not adhering to regulations. Second, there is the risk of losing competitive advantage due to data leaks. Third, as a side effect of the two previous risks, security incidents might harm an organization's reputation.

A data leak means that sensitive data, such as customer lists or cost calculations, leave the company. Other examples are engineering drawings stored in CAD systems, research data in pharmaceutical companies, or source code in the software industry. If companies lose such data to competitors, this threatens their position in the market.

The focus of data-related regulatory risks is customer data. The risks correlate especially with a worldwide customer base, outsourcing, or global work distribution. There are standards such as the Payment Card Industry Data Security Standard (PCI-DSS) or the Health Insurance Portability and Accountability Act (HIPAA). There are European or Swiss data protection laws and the EU-US safe harbor agreement. They impact whether data can be transferred to a subsidiary or to sourcing partners in the same or in a different jurisdiction. Violating any of the regulations can harm the reputation and result in interventions of regulatory bodies and fines. When employees violate laws, even if instructed to do so by their superiors, there is also a direct personal risk for them.

Risks in Development, Test, and Production Environments

Even if systems are engineered and operated securely, and IT and business enforce the need-to-know principle with roles and a strict user management, the data-related risks remain. Their root cause is normal users using their normal access rights, just not as intended. Table 1 matches abstract business risks with IT security incidents, for which concrete solutions can be defined.

The first business risk is that sensitive data leave the company. This can happen by mistake. For example, a user sends an email to a wrong person or attaches a wrong file. There is also the risk of transferring data outside of the company as part of industrial espionage, e.g., by sending data to a personal account or copying it to a USB stick. Business users working in production are the source of such risks as are engineers in development and test. The latter can

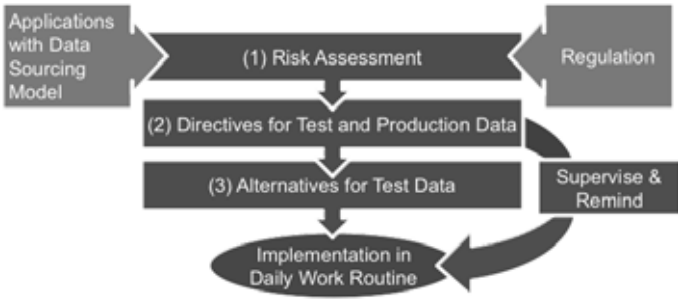

Figure 1: The three preparation steps for effective data-related risk mitigation

be even a higher risk. In production, the need-to-know principle is often enforced strictly. This reduces the number of persons who can misuse (large amounts of) sensitive data. Also, direct database access is limited to the small group of admins. In development and test environments, by contrast, engineers often have access to applications without any authentication. To make things worse, they can connect to the database directly and submit SQL queries. If a test database contains production data, engineers can extract, for example, a complete customer list with one single query. Still, there is a reason why many test environments contain production data: engineers need appropriate test data and database copies are a convenient solution.

The risk of violating regulations is obvious in production environments. A server with customer data must be placed only in datacenters in an appropriate jurisdiction. Centralized server provisioning reduces the error risk in production. The same risk exists in development and test environments, however, for which outsourcing and offshoring is much more common. Here, decisions are often made in a decentralized way. This increases the risk that sensitive production data gets into offshored or outsourced development and test environments via a database copy, file transfers, or manually entered data.

The regulatory risks increase if a company's business spreads across various countries. Consolidated datacenters and centers of excellence for certain areas (e.g., payroll processing) reduce costs; however, the more production data are transferred around the globe, the higher the risk of violating regulations.

Aimless Activism vs. Effective Risk Mitigation

When companies and risk managers understand the data-related risks, some managers might think about writing an email including the following three policies:

- Customer data must remain within our country!
- Sensitive data must not be copied into test environments!
- Intellectual property and customer data must not leave the company!

Such an email may increase awareness, but most of all, it causes confusion. These policy statements are ambiguous. On one hand, it is unclear what exactly is prohibited. "Sensitive data" is a broad term. On the other hand, developers and testers do not know what they should do instead. They rely on adequate data for their work. Thus, before sending such emails, managers should go through three preparation steps (see Figure 1). In the first step, the legal and the IT risk departments together assess the risk. Which data are sensitive from a business and a regulatory point of view? The outcome is a list of the risks with the severity of a potential incident and the probability of an occurrence.

The second step is to elaborate a directive for production and test data. The management must decide which risks it accepts. The directive must provide a data classification scheme, which explains in detail which data is defined to be sensitive. It must identify suitable datacenter locations and state whether outsourcing is possible and to which partners and jurisdictions. The directive should also define for development and test environments which data can be transferred to whom, in which jurisdiction testing is allowed, and details regarding test data anonymization (if applicable). What must be anonymized? Is it sufficient to delete the customer names only? Are customer addresses sensitive as well? What about booking texts or contracts with suppliers?

Such a production and test data directive restricts the work of developers and testers. Thus, the third step is about providing alternatives. Synthetically generated test data or (very good!) anonymized data from production environments could replace the complete database copies from production to development and test environments (see [1] for details). These alternatives might require new tools, can change the organization and its teams and processes, and, thus, can tie up resources and take time.

Business Risk	Production	Development and Test
Competitive advantage	Data loss by mistake	Data loss by mistake
	Data loss due to criminal act	Date loss to criminal act
Regulatory	Datacenter / production servers placed in inappropriate jurisdiction	Data transferred to / typed in an inappropriate jurisdiction or sourcing partner

Table 1: Risk types and environments

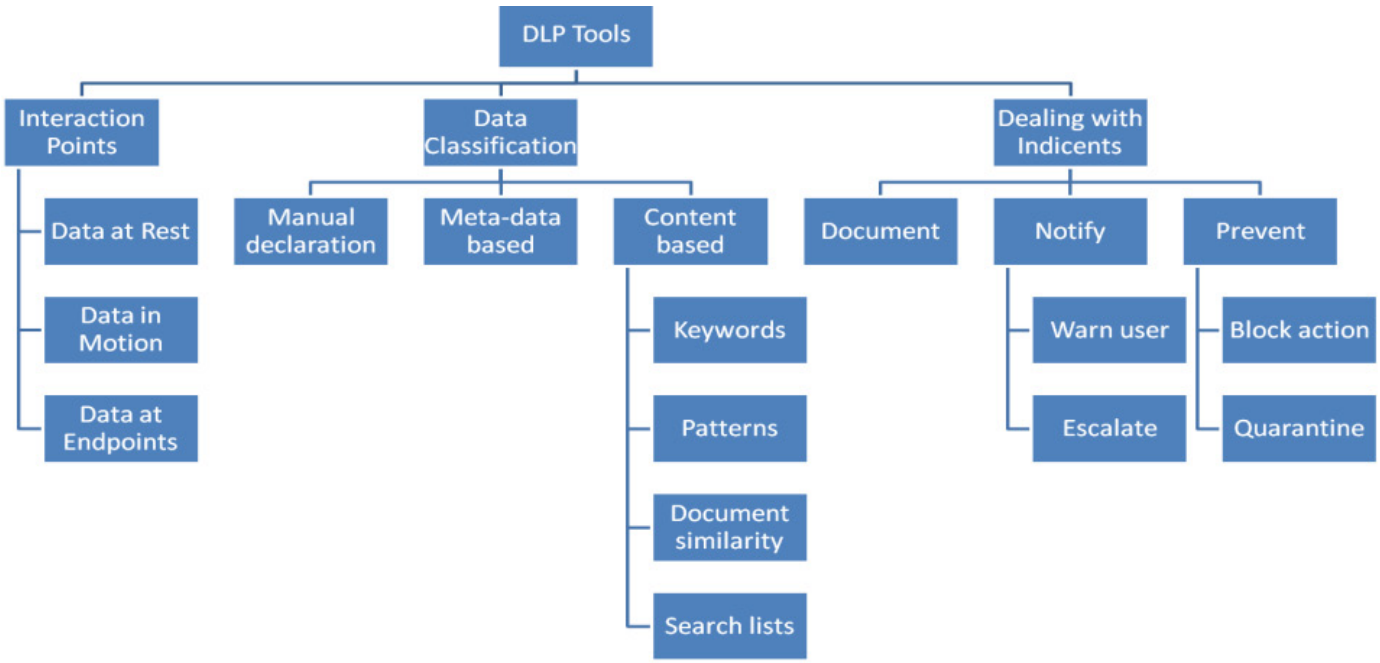

Figure 2: Characterizing DLP tools

When all three steps are completed, the management communicates the directive together with the alternatives for testers and developers. From this moment on, users in production as well as testers and developers must follow the directive; however, IT environments are similar to teenagers' rooms. Forcing them to clean up once does not ensure that everything is perfect for the next weeks. Regular checks for sensitive data are required, for which data loss prevention (DLP) tools can help.

Helpful DLP Tools

Various vendors offer data loss prevention tools. The market is dynamic and features vary. There are comprehensive solutions from the big players, such as McAfee and Symantec, or from smaller vendors, such as myDLP. Others focus on niches (e.g., Proofpoint or Microsoft Exchange). Three questions help to characterize a product or a concrete installation (Figure 2):

1. What do DLP tools look at (interaction points)?
2. How do they identify sensitive data?
3. What options are provided to react to incidents?

Three options exist for the interaction points between the DLP tool and the IT infrastructure (Figure 3):

- Data at Rest. The DLP tool searches for sensitive data in files, SharePoint servers, databases, or other kinds of repositories. The idea is to find sensitive data at places that nobody is aware of. Certainly, enterprise resource planning systems store sensitive data, e.g., customers, costs, and profits. But in many companies, critical data exist in many files as well—e.g., Excel spreadsheets.
- Data in Motion. Data are transferred within the company and to outside recipients via the network, e.g., by emails, FTP, or social media. The DLP solution can monitor the zone-internal network traffic for sensitive data as well as the traffic to other zones or the Web.
- Data at Endpoints. Here, laptops, PCs, and mobile devices are the focus. They can get lost with data on them or data can be copied from them to removable devices such as USB sticks. So it is desirable to monitor data downloads as well as data-related activities on endpoints.

When DLP tools detect sensitive data in an email or in a user's spreadsheet, they must react. Standard options are:

- Log and document the security incident in the DLP tool or in a central information security management system (ISMS).
- Notify users when they try to perform noncompliant actions, or escalate such incidents to their line manager or the HR department.
- Prevent wrongdoing by blocking actions (e.g., emails for data in motion, or downloads for data at endpoints) or quarantine files by moving them to a secure folder (data at rest).

Blocking wrongdoing seems to be the best idea, but it is not always true. If the DLP tool blocks half of the employees' emails "to be on the safe side," the DLP tool will be switched off within minutes. Thus, a first phase is always about improving the rules for data classification. But even afterwards, notifying the user or writing logs and evaluating them periodically remains the option-of-choice for less severe incidents.

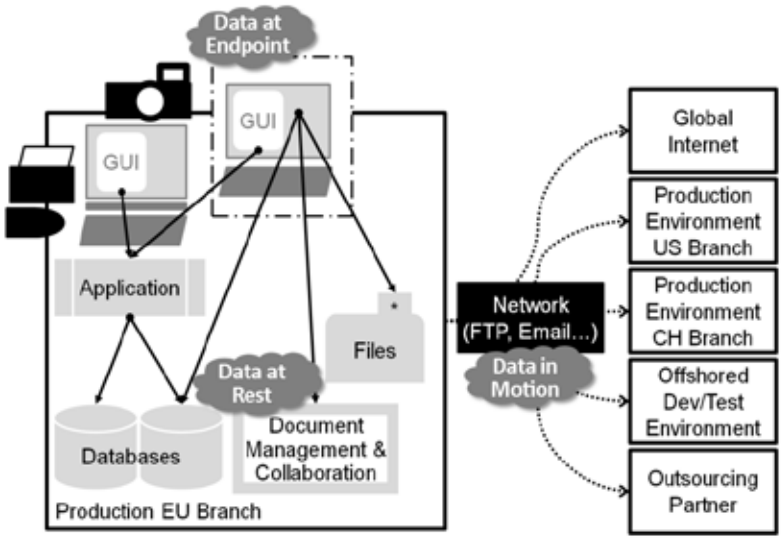

Figure 3: Interaction points in a multinational company with various network zones

The biggest challenge is data classification. The DLP tool must decide whether data, emails, files, etc. are sensitive. Options are:

- Manual declaration. A risk manager tags a folder or files as sensitive. From this moment on, the DLP tool prevents screen dumps, for example, when a sensitive file is shown on the screen.
- Content-based classification. The DLP tool looks inside files or emails. The main techniques are (1) keyword search, (2) document similarity, (3) patterns, and (4) search lists. Keywords are strings whose appearance signal sensitivity to the DLP tool, e.g., a term "strictly confidential." Document similarity means that the DLP tool has a collection of sensitive files, for example, templates for offers or contracts. Similar files are assumed to be sensitive. So the DLP could be triggered if a user tries to send out hundreds of contracts. Identifiers such as credit card numbers or social security numbers often have a specific format, e.g., four digits, a space, four digits. Patterns allow searching for such identifiers in emails or files. Finally, search lists provide a list of sensitive data items, such as all customer email addresses or customer credit cards. If one item appears in a file or email, for example, the DLP tool raises an incident.
- Metadata-based classification (e.g., names or IP address ranges) helps when deciding about sensitiveness. They can be combined with content-based strategies. Then, emails with sensitive data can be sent within the company, but the DLP tool prevents such emails from being sent out.

The big challenge is to configure the DLP tool such that it finds "real" incidents without raising many false alarms. Database/SQL developers might help more than security consultants with a background in firewalls and virus scanning.

DLP Features and Risk Reduction

The DLP tool can reduce the risk of criminal or accidental disclosure of sensitive data with its data-in-motion and data-at-endpoint features (see Table 2). They identify and block such data transfers via the network or to mobile devices and USB sticks. The data-in-motion features monitor data transfers by email or file to other jurisdictions, to sourcing partners, or to development environments. Database transfers can be tricky for DLP tools. In this case, the IT department's database copy process must prevent inappropriate data transfers. Still, DLP tools can help in assessing whether a database contains sensitive data. As stated in the example with the teenager's room, periodic checks are needed, especially in test and development environments.

The data-at-rest features can reduce the overall exposure to data-related risks. Periodic sanity checks of file systems or SharePoint servers can find unofficial data collections. Cleaning them up means less sensitive data will be floating around. This

Concrete risk	How DLP helps	Data at/in...		
		Rest	Motion	Endpoints
Data loss by mistake Data loss due to criminal act	Inappropriate data transfers (e.g., email, FTP, mobile devices, or USB sticks) identified and blocked		X	X
Datacenter/production servers placed in inappropriate jurisdiction	n/a			
Data transferred to/held in inappropriate jurisdiction	Entry check during file/database transfers	(X)	X	X
	Periodic sanity check of files/databases	X		
(Spread of sensitive data)	Periodic sanity check of files	X		

Table 2: How DLP helps reduce risk

lessens the risk that such data get lost or transferred to a wrong jurisdiction, environment, or outsourcing partner.

Limitations and Risks

DLP tools are 100% reliable when searching for a 30-chars long, alphanumeric string such as UAW-47594W48406DE488242O34333W. Searching for all emails or documents about a customer or patient is much more difficult. If the name is "Peter James Miller," how might the contacted person react to reading the salutation "Dear Mr. Miller"? Or to "Peter Miller" or "Dear James Miller"? And how do you ensure that you do not confuse "Peter James Miller" with a non-sensitive name "Peter Max Miller" or "Peter Miller"? Similar to any information retrieval system, the DLP tool must balance the risk of not finding certain incidents and the risk of raising too many false incidents. Important terms are recall (if there are 100 incidents that should be found, how many will you find?) and precision (out of 100 incidents raised, how many are true incidents?). It means balancing the risk of leaking important data and violating laws against having too many incidents, which cannot be handled. No company can afford to have an IT security officer read every second email, not to mention the impact on the work environment. One sub-problem is format issues. So what happens if a social security number "123-45-6789" is written as "123456789" or "123 45 6789"? Companies can enforce standards for the data in databases, but this is nearly impossible for emails and Excel or Word documents.

Besides the limitations, there are several risks: laziness, non-adequacy, circumvention, and data loss of the DLP tool. Laziness reflects that users start relying on the DLP tool instead of thinking for themselves about what is allowed. This is dangerous because DLP tools cannot find all critical events. Non-adequacy means using a DLP tool to clean up files and data. DLP tools are good at detecting a broad variety of violations, but when DLP tools are used to clean up exactly the data items the DLP tool finds, there is a nearly 100% probability that sensitive data remains, which the DLP tool did not and will never find. Only a root cause analysis of the incidents leads to an understanding where and why sensitive data shows up at the wrong places. Circumvention means that users, especially with criminal intentions, search for ways to fool the DLP tool when they learn how the DLP tool works in detail.

Finally, the biggest risk can be the DLP tool itself. If it stores customer lists or sensitive documents to find copies or similar documents, losing data from the DLP tool becomes the worst case scenario. This includes, first, the direct loss of unencrypted, sensitive data such as customer or credit card lists. Second, there is the risk of telephone book attacks. For example, the DLP tool might be directed to a large list of potential client names, creating an incident for each real client name. So the set of incidents is the full client list. Third, even if lists and documents are encrypted or hashed, they must be highly protected and must never end up on mobile devices. If the DLP tool is not open source, it is never clear how strong the encryption is and whether attackers related to governmental agencies have back doors to break the encryption.

In conclusion, data loss prevention tools enable companies to detect and prevent inappropriate data handling. This allows companies to address regulatory risks and risks related to the loss of intellectual property.

References

[1] K. Haller, "Test Data Management in Practice: Problems, Concepts, and the Swisscom Test Data Organizer," Software Quality Days 2013, Vienna, Austria.