

Chapter 2

Data Leakage

Data leakage is defined as the *accidental or unintentional distribution of private or sensitive data to an unauthorized entity*. Sensitive data in companies and organizations include intellectual property (IP), financial information, patient information, personal credit-card data, and other information depending on the business and the industry. Data leakage poses a serious issue for companies as the number of incidents and the cost to those experiencing them continue to increase. Data leakage is enhanced by the fact that transmitted data (both inbound and outbound), including emails, instant messaging, website forms, and file transfers among others, are largely unregulated and unmonitored on their way to their destinations. Furthermore, in many cases, sensitive data are shared among various stakeholders such as employees working from outside the organization's premises (e.g., on laptops), business partners, and customers. This increases the risk that confidential information will fall into unauthorized hands. Whether caused by malicious intent or an inadvertent mistake by an insider or outsider, exposure of sensitive information can seriously hurt an organization. The potential damage and adverse consequences of a data leakage incident can be classified into two categories: direct and indirect losses. Direct losses refer to tangible damage that is easy to measure or to estimate quantitatively. Indirect losses, on the other hand, are much harder to quantify and have a much broader impact in terms of cost, place, and time [Bunker, 2009]. Direct losses include violations of regulations (such as those protecting customer privacy) resulting in fines, settlements or customer compensation fees; litigation involving lawsuits; loss of future sales; costs of investigation and remedial or restoration fees. Indirect losses include reduced share price as a result of negative publicity; damage to a company's goodwill and reputation; customer abandonment; and exposure of intellectual property (business plans, code, financial reports, and meeting agendas) to competitors.

Data leakage can occur in many forms and in any place. In a 2009 Data Breach Investigation Report¹ (by the Verizon Business RISK team), 90 data breaches occurring in 2008 were analyzed. In addition to the significant number of compromised records (285 million), the investigation revealed other interesting aspects of this problem as well. One of the most intriguing aspects revealed by the compiled data is that most breaches have been caused by external parties (74%). However, the number of breaches resulting exclusively from the actions of insiders is still significant (20%). Incidents in which business partners have been involved account for 32% of the total. According to the nonprofit consumer organization Privacy Rights Clearinghouse,² a total of 227,052,199 individual records containing sensitive personal information were involved in security breaches in the United States between January 2005 and May 2008.

Some recent high-profile leakage incidents, selected from www.datalossdb.org, are presented in Table 2.1. This sample of recent leakage incidents emphasizes the difficulty of providing a “one-stop-shop” silver-bullet solution for preventing all data leakage scenarios. The sample also indicates that enterprises should broaden the focus of their security efforts beyond merely securing network perimeters and internal hosts from classic threats (i.e., viruses, Trojan horses, worms, D/DoS attacks and intrusions). In addition, organizations are obligated to comply with federal and state regulations which aim to protecting financial and other private data by directing organizations to protect their networks and data. Examples of such regulations are the Health Insurance Portability and Accountability Act (HIPAA), the Gramm-Leach-Bliley Act (GLBA), California’s data-breach disclosure notification law SB 1386, the Payment Card Industry Data Security Standard (PCI-DSS) and the Sarbanes–Oxley Act (SOX) [Frost & Sullivan, 2008].

In fact, according to the Gartner report [Ouellet, 2009], large enterprises already understand the need to use data leakage prevention (DLP) technology as one component in a comprehensive plan for the handling and transmission of sensitive data [Ouellet, 2009]. The technological means employed for enhancing DLP can be divided into the following categories (Figure 2.1): standard security measures, advanced/intelligent security measures, access control and encryption, and designated DLP systems [Phua, 2009].

Standard security measures are used by many organizations and include common mechanisms such as firewalls, intrusion detection systems (IDSs), and antivirus software that can provide protection against both outsider attacks (e.g., a firewall which limits access to the internal network and an intrusion detection system which detects attempted intrusions) and inside attacks (e.g., antivirus scans to detect a Trojan horse that may be installed on a PC to send confidential information). Another example is the use of thin clients which operate in a client-server architecture, with no personal or sensitive data stored on a client’s computer. Policies and training for improving the awareness of employees and partners provide additional standard security measures.

¹ http://www.verizonbusiness.com/resources/security/reports/2009_databreach_rp.pdf

² <http://www.privacyrights.org/ar/ChronDataBreaches.htm>

Table 2.1 Data leakage incidents

Date	Organization	Description
Oct. 2008	UPS	A UPS employee's laptop containing payroll information for 9000 U.K. employees was stolen. In response UPS announced that it will encrypt all data stored on all the company's mobile devices.
Sept. 2011	Science Applications International Corp	Backup tapes stolen from a car containing 5,117,799 patients' names, phone numbers, Social Security numbers, and medical information.
Oct. 2009	U.S. National Archive	U.S. National Archive and Records administration improperly disposed of hard drives containing 76 million names, addresses, and SSNs of US military veterans.
July 2008	Google	Data were stolen, not from Google offices, but from the headquarters of an HR outsourcing company, Colt Express. The thieves broke in and stole company computers containing unencrypted data including names, addresses and SSNs of Google employees. As a result, Google terminated its partnership with Colt Express.
Jan. 2008	Stockport Primary Care Trust (U.K.)	A member of staff lost a USB memory stick containing data extracted from the medical records of patients. The data were being carried personally to avoid sending them by e-mail because the employee thought that they would be more secure.
June 2004	AOL	An employee of America Online Inc. stole the computerized employee identification code of another AOL worker to gain access to AOL's subscriber data. He then stole 92 million email addresses belonging to 30 million subscribers and sold them to spammers.
July 2009	American Express	DBA stole a laptop containing thousands of American Express card numbers. The DBA reported it stolen, "...he (DBA) was one of the few who could have possibly downloaded all their account holders' information, including the PIN numbers used to access money from ATM machines at various banks."
2007	Wagner Resource Group	An employee of a McLean investment firm decided to trade some music using a file-sharing network while using the company computer. In doing so, he inadvertently opened the private files of his firm, Wagner Resource Group, to the public. Social Security numbers, dates of birth, and names of 2,000 clients were exposed.
Aug. 2007	Nuclear Laboratory in Los Alamos	An employee of the U.S. nuclear laboratory in Los Alamos transmitted confidential information by email. The incident was classified as a serious threat to the country's nuclear safety.

(continued)

Table 2.1 (continued)

Date	Organization	Description
Feb. 2008	Eli Lilly & Co.	One of Eli Lilly & Co.'s subcontracted lawyers at Philadelphia-based Pepper Hamilton mistakenly emailed confidential Eli Lilly discussions to <i>Times</i> reporter Alex Berenson (instead of to Bradford Berenson, her co-counsel), costing Eli Lilly nearly \$1 billion.
Sep. 2007	Scarborough & Tweed	The Web servers of Scarborough & Tweed, a company that sells corporate gifts online, were compromised and information about 570 customers may have been accessed using an SQL injection attack. The information included customers' names, addresses, telephone numbers, account numbers, and credit card numbers.
May 2009	Alberta Health Services	Personal health information on thousands of Albertans was skimmed from the Alberta Health Services Edmonton network as a computer virus infected the network and stole medical information on 1,582 people, including laboratory test results and diagnostic imaging reports. The virus captured information from a computer screen and then transmitted it to an external website.
Apr. 2009	Prague hotel (Czech Republic)	A data leakage incident occurred in a Prague hotel (Czech Republic). The flight details and passport numbers of approximately 200 EU leaders were leaked by accident. The data was related to an EU-US summit held in Prague and attended by U.S. President Obama.
Jan. 2009	Heartland Payment Systems	Malicious software/hack compromised tens of millions of credit and debit card transactions. "The data include the digital information encoded onto the magnetic stripe ... thieves can fashion counterfeit credit cards..."
2003	British Intelligence	A British intelligence report in the form of a Word document containing the names of the authors of a paper in its revision log metadata was cited by the United States in a speech to the United Nations. The metadata showed that the report was in fact written by U.S. researchers.

Creating and enforcing organization-wide data handling policies based on industry regulations and on the organization's specific requirements is essential to regulate all aspects of handling personal data in an organization. These policies declare strict rules for handling these data, such as discarding or archiving unneeded personal data and creating access control mechanisms to enable access to such data by authorized employees only. The creation of a data handling policy should be accompanied by appropriate training that informs employees of the rules and a requirement

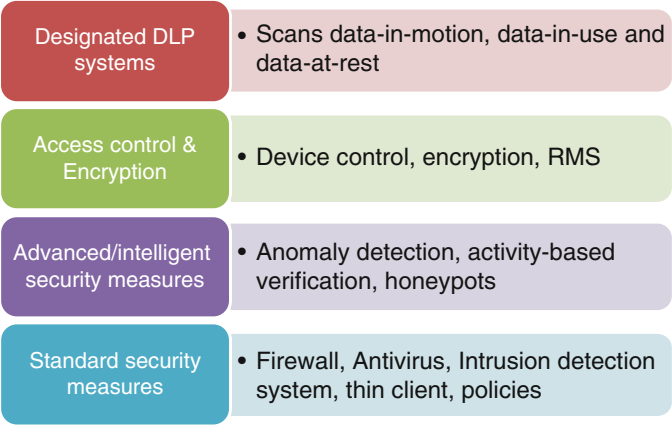


Fig. 2.1 Categories of technological approaches used to provide data leakage detection and prevention

that employees sign binding statements regarding their responsibilities and their commitment to work according to the policy.

Advanced or intelligent security measures include machine learning and temporal reasoning algorithms for detecting abnormal access to data (i.e., databases or information retrieval systems), activity-based verification (e.g., based on keystrokes and mouse patterns), detection of abnormal email exchange patterns, and applying the honeypot concept for detecting malicious insiders.

Device control, access control, and encryption are used to prevent access by an unauthorized user. These are the simplest measures that can be taken to protect large amounts of personal data against malicious outsider and insider attacks.

Designated DLP solutions are intended to detect and prevent attempts to copy or send sensitive data, intentionally or unintentionally, without authorization, mainly by personnel who are authorized to access the sensitive information. A major capability of such solutions is an ability to classify content as sensitive. Designated DLP solutions are typically implemented using mechanisms such as exact data matching, structured data fingerprinting, statistical methods (e.g., machine learning), rule and regular expression matching, published lexicons, conceptual definitions and keywords [Ouellet, 2009].

This survey focuses mainly on the category of designated Data Leakage Prevention (DLP) solutions, often referred to as *Information Leak Prevention (ILP)*, *Data Leak/Loss Prevention (DLP)*, *Outbound Content Compliance*, *Content Monitoring and Filtering*, *Content Monitoring and Protection (CMP)*, or *Extrusion Prevention* [Mogull, 2007].

Several definitions have been proposed for describing designated DLP solutions. Frost & Sullivan (2008) defined a DLP solution as a “system that monitors and enforces policies on fingerprinted data that are at-rest (i.e., in storage), in-motion (i.e., across a network) or in-use (i.e., during an operation) on a public or private

computer/network.” The report claims that ideal DLP solutions should provide data protection at the gateway and the endpoint using data discovery, which tags and fingerprints sensitive data. The tagging and fingerprinting of data will assist in enforcing policies, regulations, and laws as required by the organization. Ouellet and Proctor (2009) uses the term “content-aware DLP” to refer to a set of inspection techniques used to classify data while at-rest, in-use, or in-motion and to apply pre-defined policies (for example, logging, reporting, relocating, tagging, or encrypting). Mogull (2007) defines DLP solutions as systems that identify, monitor, and protect data-in-use, data-in-motion, and data-at-rest through deep content inspection using a centralized management framework. In this work, a designated data leakage prevention solution is defined as *a system that is designed to detect and prevent the unauthorized access, use, or transmission of confidential information.*

This book presents a methodical description of state-of-the-art research and of existing commercial DLP solutions. In contrast to the work of Hackle and Hauer (2009) who have focused on the domain of commercial DLP products, here both commercial solutions and academic research will be discussed and analyzed. To the best of the authors’ knowledge, this survey is the first to provide a review and discussion of both research and existing commercial DLP solutions. A taxonomy of DLP solutions and a classification of the security measures used for DLP will first be presented. Second, the main data leakage scenarios will be described, with for each scenario, the most relevant and applicable solution or approach that will mitigate and reduce the likelihood and impact of the leakage scenario.

A Survey of Data Leakage Detection and Prevention
Solutions

Shabtai, A.; Elovici, Y.; Rokach, L.

2012, VIII, 92 p. 9 illus., Softcover

ISBN: 978-1-4614-2052-1