SECURITY

Data Leak Prevention: Coming Soon To A Business Near You

Robin Layland

Keep critical customer data from dripping out of your hands.

Today's headlines talk about credit card numbers being stolen, customer information leaked and medical records being given out to the wrong people. It is embarrassing and it could cost companies a lot of money. Laws and regulations such as HIPAA (Health Insurance Portability and Accountability Act), Gramm-Leach-Bliley and numerous state laws are requiring that customers be notified when their personal information is stolen or lost. Sarbanes-Oxley requires that management implement internal controls over financial information (see *BCR*, April 2007, pp. 26–31).

The cost of notification and cleaning up from a data leakage can range as high as $150–$200 per customer. A "small" loss of 10,000 card numbers could cost between $1.5 and $2 million. It may even get more expensive. A bill recently introduced in Massachusetts would require the company leaking personal data to assume all the cost, including any fraud-related losses and identity theft-related expenses. Even if the cost were a fraction of that $150, a few breaches would quickly add up to serious money.

The financial hit is not the only reason to try to stop data leakage. Losing confidential information could change customer behavior. If consumers stop using online services—or even cut back—because a breach scared them away from the online world, this could cost a significant amount of money. Online transaction costs are approximately one-tenth of their brick-and-mortar equivalent. If the breach causes a small percentage of customers to go back to brick-and-mortar transactions over time, this cost could be greater than the cost of leak prevention systems.

Finally, leak prevention can be thought of as insurance. Leak prevention lessens the risk of a breach. The hope, as with all insurance, is that the premium paid to install a leak prevention solution is less than the potential loss. It also has the benefit that if there were a breach, at least something was done to try to stop it. Not having a system in place means that the company didn't even try. The lack of effort makes the leaking company look uncaring, while having a system in place with an audit trail shows that the company cares enough about customer data that it tried to prevent the leak. This could be the difference between being convicted and acquitted, in both a legal sense and in the court of public opinion.

A new market segment has recently been created to address this problem. It goes by many names, including *data leakage prevention, information leak prevention, content monitoring, extrusion prevention systems* and *privacy compliance systems*. It doesn't matter what the name, which will be settled over time; the goal is to try to minimize the loss of legally protected or valuable data.

A common but incorrect reaction is that the issue primarily affects the health care and financial industries. Most of the existing laws and regulations have indeed concentrated on medical and financial information, and both those industries have been the early adopters, but no one should think the problem is limited to just those segments. Every company has employee records that include Social Security numbers, or customer records that contain credit card information and other data from their customers that, if lost, would cause legal problems and upset customers. Additionally, losing sensitive information such as business plans, customer lists and salaries would cause problems if the information got into the wrong hands.

What It Does

Data leak prevention can do a lot to reduce leakages, but it cannot stop every leakage. A smart and determined outsider can find ways around the tools. For example, you can hide a Social Security number by putting text between the individual numbers—but the bad guy would know how to reverse the text out of the number. As a general rule, data leakage tools do not stop sophisticated criminals, but they can stop the less professional thief, something that is still very important.

The tools additionally don't stop all the *internal* bad guys. If a person has access to sensitive

Robin Layland, president of Layland Consulting, is a leading industry analyst with many years experience working for leading enterprises. He specializes in new networking technology with his current focus on critical application delivery issues including application acceleration, data center transformation and application security issues. He can be reached at Robin@Layland.com or at 860-561-4425.

data, then there is only so much that the tools will be able to do to stop them from taking the data. For example, if a customer service representative sees customer data, they can write the data down and walk out with it.

The good news is that the new tools can stop the person from sending the data out electronically, writing it to a USB drive or sending it to a printer. Data leakage tools can make it harder to steal data, especially large amounts of it; but these tools can't eliminate the risk altogether. Stopping the determined internal thief is still a task for traditional human resource methods and law enforcement.

The tools also can stop the most common leakages—unintentional leaks. Data leakage tools excel at stopping unintentional leaks. They stop the internal users from sending data to someone that shouldn't get it or copying it to a USB drive when they shouldn't be making copies. For example, a customer service representative is following up on a problem and includes sensitive information in an emailing to an unauthorized person. They are not thinking about the sensitive information and mean no harm, but if the email got out it could cause problems. The tools can stop this type of leakage.

Some systems create "fingerprints" of the data

Data leakage systems can protect sensitive information and go a long way to solving the problem, but that does not mean they can be used to protect everything. Protection is always a balancing act. There needs to be a balance between protecting valuable data while allowing the legitimate sharing and use of the information. Put too many restrictions on accessing data, and employees may have trouble performing their jobs. Tell the system to try to catch too much and the false-positive rate becomes unacceptable. Data leakage systems are also not a replacement for access control; they are a supplement, trying to catch an authorized user from sending it to an unauthorized person.

What Is Sensitive Data?

The first step in implementing a data leakage solution is to determine what needs to be protected and then look for exact matches. The most basic level is protecting data, such as credit card numbers and medical information, that falls under government-mandated regulations. Many vendors make this task easy by automatically implementing dictionaries of common terms and keywords, and can search for items like credit card numbers, medical codes and Social Security numbers.

They also allow businesses to create their own list of keywords. For example, if there is a secret project within the company, such as takeover plans, the data leak solutions can be configured with a list of terms associated with the project and then search for those items in documents.

The solutions also allow any content—databases, documents and any other content—to be marked as protected. One data leakage solution vendor, Palisade, has teamed with a music industry company called Audible Magic to incorporate the latter's database of copyrighted music. While this is not sensitive content that leaves the corporation, the system does prevent the corporation from being liable from downloading someone else's copyrighted material.

There are two big issues in protecting the data. The first is to limit the number of false positives. If the system "cries wolf" too often when there is no sensitive data, then the system becomes useless because security operations staff starts to ignore the alerts. Not every 9-digit number is a Social Security number.

A common way to reduce false positive is to look at the context. If the system sees what it thinks is a credit card number, it then looks to see if other personal data is also present. A credit card number without a name or address is generally useless. If the system doesn't see these other items in the data, then it lets it pass. Also, if the system detects the leak as a single or just a handful of credit card numbers, for example, it can be told not to alert. Most thefts are of a large number of items, not just a single item.

The second issue is how to find sensitive data that has been changed. The sensitive data is still there, just not in its exact original form. A simple example demonstrates the problem: A sensitive business plan is marked as protected, and the data leakage system has been told to be on the lookout for copies leaving the company or going to unauthorized personnel. An executive gets a copy and makes comments throughout the document. The executive then sends it in an email to someone outside the company—something not allowed. Looking only for an exact match would not find the business plan. So how data leakage systems look for this kind of "fuzz" data is one of the major differentiators between the vendors.

In the example of the business plan, some vendors attack the problem by taking a fingerprint or hash of the original document. The fingerprint is how the tool internally represents the data. A fingerprint is used instead of an exact copy because keeping a copy would present a security problem—if someone stole the data leakage device, they would have the sensitive data, and having the device itself would allow them to efficiently perform the matching function.

Where should you monitor for leaks— at servers, PCs, gateways— or beyond?

The tools break up the document to create multiple fingerprints for a single document. For example, they would create a fingerprint for every 100 characters. Then, when the tool is doing its checking function, it can look for any of these multiple fingerprints. Thus, the business plan edited by the executive would be found because the smaller fingerprints would match the unchanged sections.

How efficiently the vendors perform the matching is another major differentiator. The tools typically allow the size of the section being fingerprinted to be configured, but managers should take the lead from the vendor on this point, since picking the right number is important. Select too small a number of fingerprints per document and the number of false positives would increase; select too large a number and altered sensitive data could be leaked, because the altered portion would be so small it might not be caught.

Managers need to decide how much or what parts of a sensitive document needs to be protected. If the executive deleted the sensitive part of the business plan and sent it, should the document still be blocked? Unfortunately these decisions are required for every piece of information that's marked Protected.

Another way data leak vendors address the fuzziness problem is by learning themes or patterns. If a company wanted to block information about a project, but only do so if other secret details are present, the data leakage system can be taught that. If the email is from marketing to an outside firm and is talking about the project without the secret formula or key concepts mentioned, then the message would not be blocked. But if it had some of the critical details, the system would stop it from being sent.

Where's Waldo?

One of the problems in building a data leakage prevention system is finding out where all the sensitive data is located. Data that started out in a database can find its way to a wide range of locations, winding up on various PCs, Excel spreadsheets and USB devices. With so many possible places for data to migrate, companies often have no idea where all the sensitive data is.

Finding out where the data resides is one of the best first uses of data leakage tools. Tools from vendors such as Tablus, GTB Technologies, Reconnex and Vontu can scan the network, searching and building a catalog of the locations of sensitive data. Even if a company has not decided to take the steps to build a leakage prevention system, the tools are very helpful in finding how big a problem there is. It also tells the organization where to direct educational efforts so that internal users will not put data at risk.

What To Look At, And Where

An important issue to ask vendors about when shopping for a data leakage solution is what formats are searched. Sensitive data can migrate to a lot of different formats, and not every tool is able to search all these different formats.

For example: A report is created from sensitive data taken from a corporate SQL database on a server located in the datacenter. The data is put in an Excel spreadsheet, and the spreadsheet is then included in a document in both Microsoft Word format and as a PDF, then emailed to various employees working on it. They revise the report and discuss it using IM, with some of the sensitive data included in the IM messages. One employee copies the report to a USB drive to take home and work on it. The finalized report is then FTPed to regional servers.

If the goal is to catch the sensitive data in any form, then the solution needs to be able to monitor all the different forms the data takes and the places it resides. This means monitoring all the forms in the above example plus other forms not mentioned, including HTTP for Web applications, CIFS file transfers, JPEGs, TIFF files, peer-to-peer applications and older client/server applications—to name just the most common.

If the solution can't handle one of the data types, then there is a hole. This does not mean a solution has to support every imaginable format. If the concern is with data leakage through email, then the tool only needs to understand the email and the various attachment formats used at the company.

The next question is where to monitor for data leaks. The complete answer is to monitor at the servers, on individual PCs and at the Internet gateways, but that does not mean everyone needs to implement the complete solution, because again the answer depends on what's the greatest threat. The top concern with many managers is stopping people from sending sensitive information from the corporate email servers. The answer is to implement a data leakage solution that has an integrated MTA (Mail Transfer Agent) or interfaces with the existing MTA. The MTA receives the email, passing it to the data leakage device which scans the email for any violations. Many vendors, including Proofpoint, Reconnex and Vericept provide email solutions in appliance and software form. Many of the email solutions can also encrypt the email and perform other functions.

That's fine for email, but the examples cited earlier show that sensitive data can leak out in many forms and places. A more general solution is data leakage appliances. The appliances can be placed in front of the various types of servers at the datacenter or at the Internet connections, checking contents in this space as it leaves. Most vendors provide such an appliance; leading vendors include Code Green, Fidelis, GTB Technologies, Intrusion, Palisade, Reconnex, Tablus, Websense, Workshare and Vontu. Vericept provides its solution as software which they also package on a Linux server.

An important deployment issue is whether the appliance should be in the direct data flow—"inline"—or off to the side, receiving a copy of the data from a tap port—the "off-line" or "out-of-band" configuration. If the appliance is only reporting violations, then managers should go with an off-line solution. Many of the vendors' appliances can be either on- or off-line, but managers should check with the vendor, as some products currently only support off-line mode.

In-line configuration is better able to perform enforcement, but there are performance issues. Port speeds are not the issue; vendors provide gigabit interfaces. The issue is how many megabits (not gigabits) the appliance can process.

First, there are no good independent performance measurements. Some vendors such as Code Green quote a 200-Mbps figure, while GTB Technologies gives a 55-Mbps throughput number—and unfortunately, it is not clear if Code Green can really handle four times more than GTB's solution. Additionally GTB offers a little better price/performance. Buyers need to dig into performance numbers the vendors offer.

Putting agents on the PC gets the user involved in protecting data

The actual throughput will vary greatly no matter what figure the vendor quotes, depending on the type of documents, the parameters (how big a segment should be examined, how many keywords and so on) and the number of sensitive items it is looking for. Even when vendors provide independent numbers to document their claims, these should be taken with a big grain of salt. The best approach is to start slowly, only asking the system to monitor for the basics, and slowly requiring it to monitor for more data, allowing the development of performance numbers for the particular installation.

The way around the performance limitation is to implement multiple appliances connected to a server load balancer, something GTB Technologies has already built into its solution. Using server load balancers has the added advantage that, if one appliance fails, the others will automatically take up the load.

Another important performance issue is the amount of delay the scanning will add. Catching all instances requires that the appliance scan the entire document. This means reassembling the document before it is forwarded, something that can add significant delays.

For many applications, such as email or even FTP, this should not be an issue, as they are store-and-forward applications. But the added delay may not be acceptable for other applications.

One way around this problem is to scan the document as it is being forwarded. This is not as bad as it initially sounds, since in many cases it only takes 40–60 percent of the document to find a problem, and once a problem is found, the appliance can disrupt the forwarding of the document. In many cases, such as a Word document or a Zip file, even if part of the document has been received, it is useless to the receiver without the entire file. The only real risk with this approach is with clear text documents, since leaking half a sensitive document can still cause problems.

Another "where" question is what ports should be examined. It is common for vendors to say they can scan all software ports, trying to differentiate themselves from vendors that have concentrated on email, Web or other common ports. While it is important that all ports be scanned, it may not be as important as the vendors imply. The network firewall should be closing down most ports anyway, and thus content flowing over is not a risk.

Plus, scanning a port is not always as useful as it sounds. The scanning is done at the TCP level; if text data is being sent over an unexpected port it can be stopped, since the data leakage device can understand it. But if an application has its own binary format, the solution may not be able to decode the format if it doesn't understand that format. The ability to understand all application formats is just as important, if not more so, than the number of ports they scan.

Protecting The PC

Once the sensitive servers and the Internet connections are protected by appliances, managers must consider protecting leakages from individual PCs. It is not enough to just stop someone from sending the data over the Internet or make sure only authorized personnel receive it. Once the data is on someone's PC, they can send it to an internal unauthorized person, print it, copy it to a CD or transfer it to a USB drive.

Many vendors provide agent solutions for the PCs. These vary from providing agents that can monitor the PC as part of an overall data leakage solution to just preventing the data from being written to device such as USB drive. General-purpose host agents are provided by many of the same companies that build appliances, such as GTB Technologies, Reconnex, Tablus, Vericept, Vontu and Workshare. More limited solutions to prevent sending sensitive data to peripherals are available from Code Green, and they plan to upgrade to a fuller solution in the near term.

The key reasons for extending protection to the individual PCs is to cut down on the alerts gener-

The system is only as good as the policies that direct it

ated by the appliance, provide a better user experience and help solve the enforcement problem. If the host agent sees a problem it can inform the user, who then can take action, such as removing the sensitive data. The agent can also encrypt sensitive material. Workshare has taken it a step further by providing the ability to convert documents to PDF format, making it harder for the receiver to transfer the information to other formats.

It is important not to underestimate the importance of having the user correct the problem. Most leakages are unintentional. Alerting the user and having them correct the problem trains the users on sensitive data policy. It also can significantly reduce the number of alerts generated by the appliances. Each data leak incidence seen by an appliance must be dealt with by security operations or the appliance. A simple example of automatic correction is for the agent to automatically encrypt sensitive data such as credit card number, but automatically correcting the problem isn't necessarily the best solution.

For example, the device would automatically delete a sensitive attachment. A better solution would be for the user to delete the sensitive sections within the document. A general rule is that the user probably knows the best answer on how to correct the problem. If it reaches security operations, then they have to spend time fixing the problem or talking with the user. It is more efficient to have the user deal with the problem and leave the really serious problems or repeat offenders to security operations.

Enforcement: Stopping The Data

When a data leakage appliance sees a violation, there are several ways it can prevent the leakage. If the data leakage device has its own MTA, it can delete the email; if it is working with the existing MTA, it can tell the MTA to do it. It is important to remember that this approach only applies to the corporate mail system and not outside Web-based mail systems that internal users may be using.

What about other types of data? Once the appliance determines sensitive data is being forwarded improperly, it drops the remaining packets in the connection or resets the connection. In-line appliances can do the dropping themselves, while off-line appliances send a reset.

A popular option for HTTP-based data is to pair the data leakage appliance with a proxy from vendors such as Blue Coat Systems, or to use a shareware proxy like Squid. The proxy can reassemble the data and pass it to the appliance. This is done using ICAP (Internet Content Adaptation Protocol) based on RFC 3507. ICAP provides a way for the proxy to pass the data to the appliance and allows the appliance to tell the proxy how to handle the message. The appliance can tell the proxy to drop the data or let it pass.

The issue of what to do when it finds sensitive data is really not as simple as just blocking it. If an exact match is found and it is going to an unauthorized outsider, the answer is easy—stop the leakage—but beyond that simple situation, there are many shades of gray. What if it is going to an unauthorized internal user? Should access be blocked or is the person just trying to do their job and the data leakage system doesn't know that? What if the data is fuzzy and not an exact match?

The problem is that the system is only as good as the policies that direct it. If the policy has not been updated for a new employee, then the system would block access. If an executive in one division asks an executive in another division to look over a business plan but the policies don't allow for cross sharing, the data leakage system will block the sharing. The executives will not understand the reason they couldn't share data. Like any policy-driven system, there needs to be an effective updating procedure.

It is common for many implementations to allow sensitive data to pass through the tool and have the tool alert or record the event. This is most likely the best way to start out. It allows the team responsible for blocking data to find out what is happening and refine their policies before they have the system act on them. Many vendors allow the tool to be taught based on the alerts, making the creation of policies easier. The alerts also allow them to know where the problems are. Most leakages are by accident or ignorance, not malicious behavior. Learning who is responsible for the leakage allows the employees to be educated which, over time, will reduce most leakages.

If the decision is made to block the data, then a procedure is needed on how to handle the blockage. It is not enough to just block the data, because if fuzzy data is being blocked, there will be false positives. A process is needed on how to quickly and effectively handle false-positives, because blocking the data is keeping someone from doing their job.

What About Encryption?

If the data is encrypted before it reaches the data leakage system, the system may not be able to detect sensitive data. This is not a problem for data leakage solutions that are integrated into the server or reside on the individual PCs, but is a problem for appliances. This is a serious issue because the amount of data being encrypted is increasing.

Appliances can deal with this in a variety of ways. The appliance can de-encrypt the data by performing a proxy function. It de-encrypts the data and examines it. If no problem was found, it forwards the encrypted version. While this is a good solution, it does involve the appliance in encryption key and certificate management, and puts an added performance burden on the device.

Another approach is to work with a device that is already performing the proxy function and can de-encrypt the data. Application acceleration and other security devices face this same problem.

Many data leakage vendors, including Websense, Reconnex and others, have elected to work with vendors such as Blue Coat to receive a de-encrypted data flow from their device. This means they are working in the off-line mode, but it does solve the problem.

Summary

What is all this protection going to cost? The low end list price for a data leakage appliance is $5,000 but most are in the $23,000 to $30,000 range. The prices go up from there with many of the vendors adopting a pricing model based on the number of users it supports. For a large installation, the cost for many vendors' solutions can climb into the $100,000 range. The prices generally include subscription and updates. The charge for a client solution is in the $30 to $60 per user range. A scanning solution that finds out where sensitive data is located throughout the network can cost in the $20,000 range.

Data leakage prevention, while new, is fast becoming a must-have for a large number of companies due to the regulatory environment and the demands of their customers. While most companies have a little time before they have to implement a solution, the truth is that most companies should start planning and learning about the market now. The field is rapidly evolving, and while there are shortcomings in all the present solutions, with this much interest in finding a better solution, most of those shortcomings will be corrected in the next few years□

Companies Mentioned In This Article

Audible Magic (www.audiblemagic.com)
Blue Coat Systems (www.bluecoatsystems.com)
Code Green (www.codegreennetworks.com)
Fidelis (www.fidelissecurity.com)
GTB Technologies (www.gttb.com)
Palisade (www.palisadesys.com)
Proofpoint (www.proofpoint.com)
Reconnex (www.reconnex.net)
Squid (www.squid-cache.org/)
Tablus (www.tablus.com)
Vericept (www.vericept.com)
Vontu (www.vontu.com)
Websense (www.websense.com)
Workshare (www.workshare.com)

Prices may range from $5,000 to around $100,000