



Securosis, L.L.C.

Understanding and Selecting a Data Loss Prevention Solution

This Report Sponsored By



Author's Note

The content in this report was developed independently of any sponsors. It is based on material originally posted on the [Securosis blog](#) but has been enhanced, reviewed by SANS, and professionally edited.

This report is sponsored by Websense Inc. and released in cooperation with the [SANS Institute](#).

Special thanks to Chris Pepper for editing and content support.

Sponsored by Websense

Websense® Content Protection Suite is an integrated data loss prevention solution that prevents data leakage – both external and internal – improves business processes, and manages compliance and risk by: discovering where data is located, monitoring it, and protecting it, securing who and what go where and how.

For more information on the Websense Content Protection Suite, visit www.websense.com/cps.

Copyright

This report is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0.

<http://creativecommons.org/licenses/by-nc-nd/3.0/us/>

Table of Contents

Introduction to DLP	5
A Confusing Market	5
Defining DLP	5
DLP Features vs. DLP Solutions	6
Content Awareness	7
Content vs. Context	7
Content Analysis	7
Content Analysis Techniques	7
Technical Architecture	11
Protecting Data In Motion, At Rest, and In Use	11
Data in Motion	11
Data at Rest	14
Data in Use	15
Central Administration, Policy Management, and Workflow	18
User Interface	18
Hierarchical Management, Directory Integration, and Role-Based Administration	19
Policy Creation and Management	19

Incident Workflow and Case Management	20
System Administration, Reporting, and Other Features	21
The DLP Selection Process	22
Define Needs and Prepare Your Organization	22
Formalize Requirements	23
Evaluate Products	23
Internal Testing	24
Conclusion	25
Navigating the Maze	25
About the Author	26
About Securosis	26
About the SANS Institute	26

Introduction to DLP

A Confusing Market

Data Loss Prevention is one of the most hyped, and least understood, tools in the security arsenal. With at least a half-dozen different names and even more technology approaches, it can be difficult to understand the ultimate value of the tools and which products best suit which environments. This report will provide the necessary background in DLP to help you understand the technology, know what to look for in a product, and find the best match for your organization.

DLP is an adolescent technology that provides significant value for those organizations that need it, despite products that may not be as mature as in other areas of IT. The market is currently dominated by startups, but large vendors have started stepping in, typically through acquisition.

The first problem in understanding DLP is figuring out what we're actually talking about. The following names are all being used to describe the same market:

- Data Loss Prevention/Protection
- Data Leak Prevention/Protection
- Information Loss Prevention/Protection
- Information Leak Prevention/Protection
- Extrusion Prevention
- Content Monitoring and Filtering
- Content Monitoring and Protection

DLP seems the most common term, and while its life is probably limited, we will use it in this report for simplicity.

Defining DLP

There is a lack of consensus on what actually comprises a DLP solution. Some people consider encryption or USB port control DLP, while others limit themselves to complete product suites. Securosis defines DLP as:

Products that, based on central policies, identify, monitor, and protect data at rest, in motion, and in use, through deep content analysis.

Thus the key defining characteristics are:

- Deep content analysis
- Central policy management
- Broad content coverage across multiple platforms and locations

DLP solutions both protect sensitive data and provide insight into the use of content within the enterprise. Few enterprises classify data beyond that which is public, and everything else. DLP helps organizations better understand their data and improved their ability to classify and manage content.

Point products may provide some DLP functionality, but tend to be more limited in either their coverage (network only or endpoint only) or content analysis capabilities. This report will focus on comprehensive DLP suites, but some organizations may find that a point solution is able to meet their needs.

DLP Features vs. DLP Solutions

The DLP market is also split between DLP as a feature, and DLP as a solution. A number of products, particularly email security solutions, provide basic DLP functions, but aren't complete DLP solutions. The difference is:

- A *DLP Product* includes centralized management, policy creation, and enforcement workflow, dedicated to the monitoring and protection of content and data. The user interface and functionality are dedicated to solving the business and technical problems of protecting content through content awareness.
- *DLP Features* include some of the detection and enforcement capabilities of DLP products, but are not dedicated to the task of protecting content and data.

This distinction is important because DLP products solve a specific business problem that may or may not be managed by the same business unit or administrator responsible for other security functions. We often see non-technical users such as legal or compliance officers responsible for the protection of content. Even human resources is often involved with the disposition of DLP alerts. Some organizations find that the DLP policies themselves are highly sensitive or need to be managed by business unit leaders outside of security, which also may argue for a dedicated solution. Because DLP is dedicated to a clear business problem (protect my content) that is differentiated from other security problems (protect my PC or protect my network) most of you should look for dedicated DLP solutions.

This doesn't mean that DLP as a feature won't be the right solution for you, especially in smaller organizations. It also doesn't mean that you won't buy a suite that includes DLP, as long as the DLP management is separate and dedicated to DLP. We'll be seeing more and more suites as large vendors enter the space, and it often makes sense to run DLP analysis or enforcement within another product, but the central policy creation, management, and workflow should be dedicated to the DLP problem and isolated from other security functions.

The last thing to remember about DLP is that it is highly effective against bad business processes (FTP exchange of unencrypted medical records with your insurance company, for example) and mistakes. While DLP offers some protection against malicious activity, we're at least a few years away from these tools protecting against knowledgeable attackers.

Content Awareness

Content vs. Context

We need to distinguish content from context. One of the defining characteristics of DLP solutions is their *content awareness*. This is the ability of products to analyze deep content using a variety of techniques, and is very different from analyzing context. It's easiest to think of content as a letter, and context as the envelope and environment around it. Context includes things like source, destination, size, recipients, sender, header information, metadata, time, format, and anything else short of the content of the letter itself. Context is highly useful and any DLP solution should include contextual analysis as part of an overall solution.

A more advanced version of contextual analysis is *business context analysis*, which involves deeper analysis of the content, its environment at the time of analysis, and the use of the content at that time.

Content awareness involves peering inside containers and analyzing the content itself. The advantage of content awareness is that while we use context, we're not restricted by it. If I want to protect a piece of sensitive data I want to protect it everywhere — not just in obviously sensitive containers. I'm protecting the data, not the envelope, so it makes a lot more sense to open the letter, read it, and decide how to treat it. This is more difficult and time consuming than basic contextual analysis and is the defining characteristic of DLP solutions.

Content Analysis

The first step in content analysis is capturing the envelope and opening it. The engine then needs to parse the context (we'll need that for the analysis) and dig into it. For a plain text email this is easy, but when you want to look inside binary files it gets a little more complicated. All DLP solutions solve this using *file cracking*. File cracking is the technology used to read and understand the file, even if the content is buried multiple levels down. For example, it's not unusual for the cracker to read an Excel spreadsheet embedded in a Word file that's zipped. The product needs to unzip the file, read the Word doc, analyze it, find the Excel data, read that, and analyze it. Other situations get far more complex, like a .pdf embedded in a CAD file. Many of the products on the market today support around 300 file types, embedded content, multiple languages, double byte character sets for Asian languages, and pulling plain text from unidentified file types. Quite a few use the Autonomy or Verity content engines to help with file cracking, but all the serious tools have quite a bit of proprietary capability, in addition to the embedded content engine. Some tools support analysis of encrypted data if enterprise encryption is used with recovery keys, and most tools can identify standard encryption and use that as a contextual rule to block/quarantine content.

Content Analysis Techniques

Once the content is accessed, there are seven major analysis techniques used to find policy violations, each with its own strengths and weaknesses.

1. Rule-Based/Regular Expressions: This is the most common analysis technique available in both DLP products and other tools with DLP features. It analyzes the content for specific rules — such as 16 digit numbers that meet credit card checksum requirements, medical billing codes, or other textual analyses. Most DLP solutions enhance basic regular expressions with their own additional analysis rules (e.g., a name in proximity to an address near a credit card number).

What it's best for: As a first-pass filter, or for detecting easily identified pieces of structured data like credit card numbers, social security numbers, and healthcare codes/records.

```
^(?:(<Visa>4\d{3})|(<Mastercard>5[1-5]\d{2})|(<Discover>6011)|(<DinersClub>3[68]\d{2})|(?  
:30[0-5]\d)|(<AmericanExpress>3[47]\d{2}))([  
-]?)(?(DinersClub)(?:\d{6}\1\d{4})|?(AmericanExpress)(?:\d{6}\1\d{5})|(?  
:\d{4}\1\d{4}\1\d{4})))$
```

Regular Expression for Credit Card Numbers

Strengths: Rules process quickly and can be easily configured. Most products ship with initial rule sets. The technology is well understood and easy to incorporate into a variety of products.

Weaknesses: Prone to high false positive rates. Offers very little protection for unstructured content like sensitive intellectual property.

2. Database Fingerprinting: Sometimes called Exact Data Matching. This technique takes either a database dump or live data (via ODBC connection) from a database and only looks for exact matches. For example, you could generate a policy to look only for credit card numbers in your customer base, thus ignoring your own employees buying online. More advanced tools look for combinations of information, such as the magic combination of first name or initial, with last name, with credit card or social security number, that triggers a California SB 1386 disclosure. Make sure you understand the performance and security implications of nightly extracts vs. live database connections.

What it's best for: Structured data from databases.

Strengths: Very low false positives (close to 0). Allows you to protect customer/sensitive data while ignoring other, similar, data used by employees (like their personal credit cards for online orders).

Weaknesses: Nightly dumps won't contain transaction data since the last extract. Live connections can affect database performance. Large databases affect product performance.

3. Exact File Matching: With this technique you take a hash of a file and monitor for any files that match that exact fingerprint. Some consider this to be a contextual analysis technique since the file contents themselves are not analyzed.

What it's best for: Media files and other binaries where textual analysis isn't necessarily possible.

Strengths: Works on any file type, low false positives with a large enough hash value (effectively none).

Weaknesses: Trivial to evade. Worthless for content that's edited, such as standard office documents and edited media files.

4. Partial Document Matching: This technique looks for a complete or partial match on protected content. Thus you could build a policy to protect a sensitive document, and the DLP solution will look for either the complete text of the document, or even excerpts as small as a few sentences. For example, you could load up a business plan for a new product and the DLP solution would alert if an employee pasted a single paragraph into an Instant Message. Most solutions are based on a technique known as cyclical hashing, where you take a hash of a portion of the content, offset a predetermined number of characters, then take another hash, and keep going until the document is completely loaded

as a series of overlapping hash values. Outbound content is run through the same hash technique, and the hash values compared for matches. Many products use cyclical hashing as a base, then add more advanced linguistic analysis.

What it's best for: Protecting sensitive documents, or similar content with text such as CAD files (with text labels) and source code. Unstructured content that's known to be sensitive.

Strengths: Ability to protect unstructured data. Generally low false positives (some vendors will say zero false positives, but any common sentence/text in a protected document can trigger alerts). Doesn't rely on complete matching of large documents; can find policy violations on even a partial match.

Weaknesses: Performance limitations on the total volume of content that can be protected. Common phrases/verbiage in a protected document may trigger false positives. Must know exactly which documents you want to protect. Trivial to avoid (ROT 1 encryption is sufficient for evasion).

5. Statistical Analysis: Use of machine learning, Bayesian analysis, and other statistical techniques to analyze a corpus of content and find policy violations in content that resembles the protected content. This category includes a wide range of statistical techniques which vary greatly in implementation and effectiveness. Some techniques are very similar to those used to block spam.

What it's best for: Unstructured content where a deterministic technique, like partial document matching, would be ineffective. For example, a repository of engineering plans that's impractical to load for partial document matching due to high volatility or massive volume.

Strengths: Can work with more nebulous content where you may not be able to isolate exact documents for matching. Can enforce policies such as "alert on anything outbound that resembles the documents in this directory".

Weaknesses: Prone to false positives and false negatives. Requires a large corpus of source content — the bigger the better.

6. Conceptual/Lexicon: This technique uses a combination of dictionaries, rules, and other analyses to protect nebulous content that resembles an "idea". It's easier to give an example — a policy that alerts on traffic that resembles insider trading, which uses key phrases, word counts, and positions to find violations. Other examples are sexual harassment, running a private business from a work account, and job hunting.

What it's best for: Completely unstructured ideas that defy simple categorization based on matching known documents, databases, or other registered sources.

Strengths: Not all corporate policies or content can be described using specific examples; Conceptual analysis can find loosely defined policy violations other techniques can't even think of monitoring for.

Weaknesses: In most cases these are not user-definable and the rule sets must be built by the DLP vendor with significant effort (costing more). Because of the loose nature of the rules, this technique is very prone to false positives and false negatives.

7. Categories: Pre-built categories with rules and dictionaries for common types of sensitive data, such as credit card numbers/PCI protection, HIPAA, etc.

What it's best for: Anything that neatly fits a provided category. Typically easy to describe content related to privacy, regulations, or industry-specific guidelines.

Securosis, L.L.C.

Strengths: Extremely simple to configure. Saves significant policy generation time. Category policies can form the basis for more advanced, enterprise specific policies. For many organizations, categories can meet a large percentage of their data protection needs.

Weaknesses: One size fits all might not work. Only good for easily categorized rules and content.

These 7 techniques form the basis for most of the DLP products on the market. Not all products include all techniques, and there can be significant differences between implementations. Most products can also chain techniques — building complex policies from combinations of content and contextual analysis techniques.

Technical Architecture

Protecting Data In Motion, At Rest, and In Use

The goal of DLP is to protect content throughout its lifecycle. In terms of DLP, this includes three major aspects:

- Data At Rest includes scanning of storage and other content repositories to identify where sensitive content is located. We call this content discovery. For example, you can use a DLP product to scan your servers and identify documents with credit card numbers. If the server isn't authorized for that kind of data, the file can be encrypted or removed, or a warning sent to the file owner.
- Data In Motion is sniffing of traffic on the network (passively or inline via proxy) to identify content being sent across specific communications channels. For example, this includes sniffing emails, instant messages, and web traffic for snippets of sensitive source code. In motion tools can often block based on central policies, depending on the type of traffic.
- Data In Use is typically addressed by endpoint solutions that monitor data as the user interacts with it. For example, they can identify when you attempt to transfer a sensitive document to a USB drive and block it (as opposed to blocking use of the USB drive entirely). Data in use tools can also detect things like copy and paste, or use of sensitive data in an unapproved application (such as someone attempting to encrypt data to sneak it past the sensors).

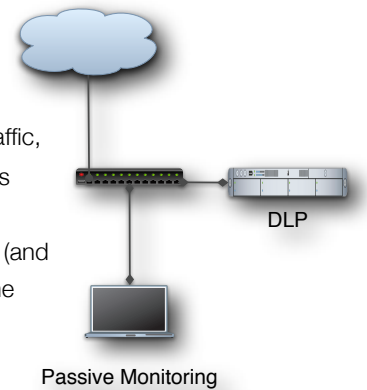
Data in Motion

Many organizations first enter the world of DLP with network based products that provide broad protection for managed and unmanaged systems. It's typically easier to start a deployment with network products to gain broad coverage quickly. Early products limited themselves to basic monitoring and alerting, but all current products include advanced capabilities to integrate with existing network infrastructure and provide protective, not just detective, controls.

Network Monitor

At the heart of most DLP solutions lies a passive network monitor. The network monitoring component is typically deployed at or near the gateway on a SPAN port (or a similar tap). It performs full packet capture, session reconstruction, and content analysis in real time. Performance is more complex and subtle than vendors normally discuss. First, on the

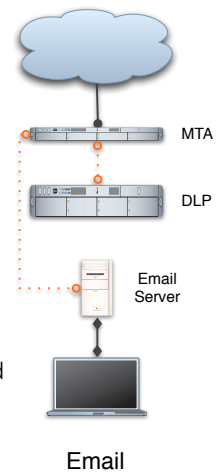
client expectation side, most clients claim they need full gigabit ethernet performance, but that level of performance is unnecessary except in very unusual circumstances since few organizations are really running that high a level of communications traffic. DLP is a tool to monitor employee communications, not web application traffic. Realistically we find that small enterprises normally run under 50 MByte/s of relevant traffic, medium enterprises run closer to 50-200 MB/s, and large enterprises around 300 MB/s (maybe as high as 500 in a few cases). Because of the content analysis overhead, not every product runs full packet capture. You might have to choose between pre-filtering (and thus missing non-standard traffic) or buying more boxes and load balancing. Also, some products lock monitoring into pre-defined port and protocol combinations, rather than using service/channel identification based on packet content. Even if full application channel identification is included, you want to make sure it's enabled. Otherwise, you might miss non-standard communications such as connecting over an unusual port. Most of the network monitors are dedicated general-purpose server hardware with DLP software installed. A few vendors deploy true specialized appliances.



While some products have their management, workflow, and reporting built into the network monitor, this is often offloaded to a separate server or appliance.

Email Integration

The next major component is email integration. Since email is store and forward you can gain a lot of capabilities, including quarantine, encryption integration, and filtering, without the same hurdles to avoid blocking synchronous traffic. Most products embed an MTA (Mail Transport Agent) into the product, allowing you to just add it as another hop in the email chain. Quite a few also integrate with some of the major existing MTAs/email security solutions directly for better performance. One weakness of this approach is it doesn't give you access to internal email. If you're on an Exchange server, internal messages never make it through the external MTA since there's no reason to send that traffic out. To monitor internal mail you'll need direct Exchange/Lotus integration, which is surprisingly rare in the market. Full integration is different from just scanning logs/libraries after the fact, which is what some companies call internal mail support. Good email integration is absolutely critical if you ever want to do any filtering, as opposed to just monitoring.



Filtering/Blocking and Proxy Integration

Nearly anyone deploying a DLP solution will eventually want to start blocking traffic. There's only so long you can take watching all your juicy sensitive data running to the nether regions of the Internet before you start taking some action. But blocking isn't the easiest thing in the world, especially since we're trying to allow good traffic, only block bad traffic, and make the decision using real-time content analysis. Email, as we just mentioned, is fairly straightforward to filter. It's not quite real-time and is proxied by its very nature. Adding one more analysis hop is a manageable problem in even the most complex environments. Outside of email most of our communications traffic is synchronous — everything runs in real time. Thus if we want to filter it we either need to bridge the traffic, proxy it, or poison it from the outside.

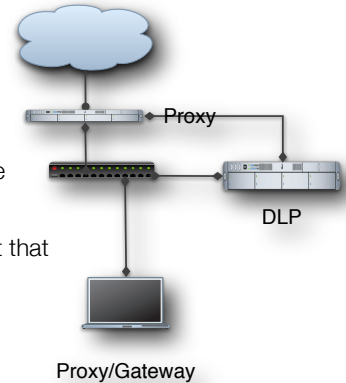
Bridge

With a bridge we just have a system with two network cards which performs content analysis in the middle. If we see something bad, the bridge breaks the connection for that session. Bridging isn't the best approach for DLP since it might not stop all the bad traffic before it leaks out. It's like sitting in a doorway watching everything go past with a magnifying

glass; by the time you get enough traffic to make an intelligent decision, you may have missed the really good stuff. Very few products take this approach, although it does have the advantage of being protocol agnostic.

Proxy

In simplified terms, a proxy is protocol/application specific and queues up traffic before passing it on, allowing for deeper analysis. We see gateway proxies mostly for HTTP, FTP, and IM protocols. Few DLP solutions include their own proxies; they tend to integrate with existing gateway/proxy vendors since most customers prefer integration with these existing tools. Integration for web gateways is typically through the iCAP protocol, allowing the proxy to grab the traffic, send it to the DLP product for analysis, and cut communications if there's a violation. This means you don't have to add another piece of hardware in front of your network traffic and the DLP vendors can avoid the difficulties of building dedicated network hardware for inline analysis. If the gateway includes a reverse SSL proxy you can also sniff SSL connections. You will need to make changes on your endpoints to deal with all the certificate alerts, but you can now peer into encrypted traffic. For Instant Messaging you'll need an IM proxy and a DLP product that specifically supports whatever IM protocol you're using.



TCP Poisoning

The last method of filtering is TCP poisoning. You monitor the traffic and when you see something bad, you inject a TCP reset packet to kill the connection. This works on every TCP protocol but isn't very efficient. For one thing, some protocols will keep trying to get the traffic through. If you TCP poison a single email message, the server will keep trying to send it for 3 days, as often as every 15 minutes. The other problem is the same as bridging — since you don't queue the traffic at all, by the time you notice something bad it might be too late. It's a good stop-gap to cover nonstandard protocols, but you'll want to proxy as much as possible.

Internal Networks

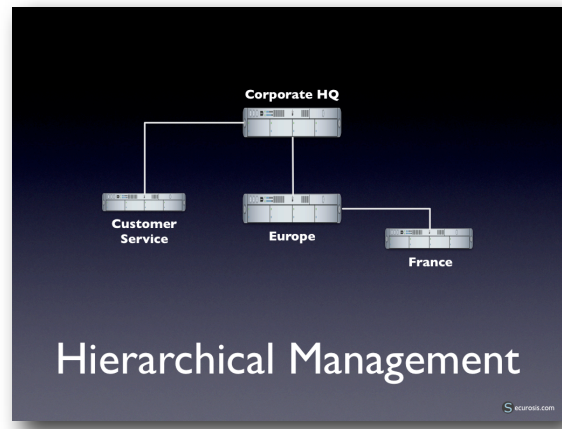
Although technically capable of monitoring internal networks, DLP is rarely used on internal traffic other than email. Gateways provide convenient choke points; internal monitoring is a daunting prospect from cost, performance, and policy management/false positive standpoints. A few DLP vendors have partnerships for internal monitoring but this is a lower priority feature for most organizations.

Distributed and Hierarchical Deployments

All medium to large enterprises, and many smaller organizations, have multiple locations and web gateways. A DLP solution should support multiple monitoring points, including a mix of passive network monitoring, proxy points, email servers, and remote locations. While processing/analysis can be offloaded to remote enforcement points, they should

send all events back to a central management server for workflow, reporting, investigations, and archiving. Remote offices are usually easy to support since you can just push policies down and reporting back, but not every product has this capability.

The more advanced products support hierarchical deployments for organizations that want to manage DLP differently in multiple geographic locations, or by business unit. International companies often need this to meet legal monitoring requirements which vary by country. Hierarchical management supports coordinated local policies and enforcement in different regions, running on their own management servers, communicating back to a central management server. Early products only supported one management server but now we have options to deal with these distributed situations, with a mix of corporate/regional/business unit policies, reporting, and workflow.



Data at Rest

While catching leaks on the network is fairly powerful, it's only one small part of the problem. Many customers are finding that it's just as valuable, if not more valuable, to figure out where all that data is stored in the first place. We call this *content discovery*. Enterprise search tools might be able to help with this, but they really aren't tuned well for this specific problem. Enterprise data classification tools can also help, but based on discussions with a number of clients they don't seem to work well for finding specific policy violations. Thus we see many clients opting to use the content discovery features of their DLP products.

The biggest advantage of content discovery in a DLP tool is that it allows you to take a single policy and apply it across data no matter where it's stored, how it's shared, or how it's used. For example, you can define a policy that requires credit card numbers to only be emailed when encrypted, never be shared via HTTP or HTTPS, only be stored on approved servers, and only be stored on workstations/laptops by employees on the accounting team. All of this can be specified in a single policy on the DLP management server.

Content discovery consists of three components:

1. Endpoint Discovery: scanning workstations and laptops for content.
2. Storage Discovery: scanning mass storage, including file servers, SAN, and NAS.
3. Server Discovery: application-specific scanning of stored data on email servers, document management systems, and databases (not currently a feature of most DLP products, but beginning to appear in some Database Activity Monitoring products).

Content Discovery Techniques

There are three basic techniques for content discovery:

1. *Remote Scanning*: a connection is made to the server or device using a file sharing or application protocol, and scanning is performed remotely. This is essentially mounting a remote drive and scanning it from a server that takes policies from, and sends results to, the central policy server. For some vendors this is an appliance, for others it's a commodity server, and for smaller deployments it's integrated into the central management server.

2. *Agent-Based Scanning*: an agent is installed on the system (server) to be scanned and scanning is performed locally. Agents are platform specific, and use local CPU cycles, but can potentially perform significantly faster than remote scanning, especially for large repositories. For endpoints, this should be a feature of the same agent used for enforcing Data In Use controls.
3. *Memory-Resident Agent Scanning*: Rather than deploying a full-time agent, a memory-resident agent is installed, performs a scan, then exits without leaving anything running or stored on the local system. This offers the performance of agent-based scanning in situations where you don't want an agent running all the time.

Any of these technologies can work for any of the modes, and enterprises will typically deploy a mix depending on policy and infrastructure requirements. We currently see technology limitations with each approach which guide deployment:

- Remote scanning can significantly increase network traffic and has performance limitations based on network bandwidth and target and scanner network performance. Some solutions can only scan gigabytes per day (sometimes hundreds, but not terabytes per day), per server based on these practical limitations, which may be inadequate for very large storage.
- Agents, temporal or permanent, are limited by processing power and memory on the target system, which often translates to restrictions on the number of policies that can be enforced, and the types of content analysis that can be used. For example, most endpoint agents are not capable of partial document matching or database fingerprinting against large data sets. This is especially true of endpoint agents which are more limited.
- Agents don't support all platforms.

Data at Rest Enforcement

Once a policy violation is discovered, the DLP tool can take a variety of actions:

- Alert/Report: create an incident in the central management server just like a network violation.
- Warn: notify the user via email that they may be in violation of policy.
- Quarantine/Notify: move the file to the central management server and leave a text file with instructions on how to request recovery of the file.
- Quarantine/Encrypt: encrypt the file in place, usually leaving a plain text file describing how to request decryption.
- Quarantine/Access Control: change access controls to restrict access to the file.
- Remove/Delete: either transfer the file to the central server without notification, or just delete it.

The combination of different deployment architectures, discovery techniques, and enforcement options creates a powerful combination for protecting data at rest and supporting compliance initiatives. For example, we're starting to see increasing deployments of CMF to support PCI compliance — more for the ability to ensure (and report) that no cardholder data is stored in violation of PCI than to protect email or web traffic.

Data in Use

DLP usually starts on the network because that's the most cost-effective way to get the broadest coverage. Network monitoring is non-intrusive (unless you have to crack SSL) and offers visibility to any system on the network, managed or unmanaged, server or workstation. Filtering is more difficult, but again still relatively straightforward on the network (especially for email) and covers all systems connected to the network. But it's clear this isn't a complete solution; it doesn't protect data when someone walks out the door with a laptop, and can't even prevent people from copying data to portable storage like USB drives. To move from a "leak prevention" solution to a "content protection" solution, products need to expand not only to stored data, but to the endpoints where data is used.

Note: Although there have been large advancements in endpoint DLP, endpoint-only solutions are not recommended for most users. As we'll discuss, they normally require compromise on the number and types of policies that can be enforced, offer limited email integration, and offer no protection for unmanaged systems. Long-term, you'll need both network and endpoint capabilities, and most of the leading network solutions are adding or already offer at least some endpoint protection.

Adding an endpoint agent to a DLP solution not only gives you the ability to discover stored content, but to potentially protect systems no longer on the network or even protect data as it's being actively used. While extremely powerful, it has been problematic to implement. Agents need to perform within the resource constraints of a standard laptop while maintaining content awareness. This can be difficult if you have large policies such as, "protect all 10 million credit card numbers from our database", as opposed to something simpler like, "protect any credit card number" that will generate false positives every time an employee visits Amazon.com.

Key Capabilities

Existing products vary widely in functionality, but we can break out three key capabilities:

1. Monitoring and enforcement within the network stack: This allows enforcement of network rules without a network appliance. The product should be able to enforce the same rules as if the system were on the managed network, as well as separate rules designed only for use on unmanaged networks.
2. Monitoring and enforcement within the system kernel: By plugging directly into the operating system kernel you can monitor user activity, such as copying and pasting sensitive content. This can also allow products to detect (and block) policy violations when the user is taking sensitive content and attempting to hide it from detection, perhaps by encrypting it or modifying source documents.
3. Monitoring and enforcement within the file system: This allows monitoring and enforcement based on where data is stored. For example, you can perform local discovery and/or restrict transfer of sensitive content to unencrypted USB devices.

These options are simplified, and most early products focus on 1 and 3 to solve the portable storage problem and protect devices on unmanaged networks. System/kernel integration is much more complex and there are a variety of approaches to gaining this functionality.

Use Cases

Endpoint DLP is evolving to support a few critical use cases:

- Enforcing network rules off the managed network, or modifying rules for more hostile networks.
- Restricting sensitive content from portable storage, including USB drives, CD/DVD drives, home storage, and devices like smartphones and PDAs.
- Restricting copy and paste of sensitive content.
- Restricting applications allowed to use sensitive content — e.g., only allowing encryption with an approved enterprise solution, not tools downloaded online that don't allow enterprise data recovery.
- Integration with Enterprise Digital Rights Management to automatically apply access control to documents based on the included content.
- Auditing use of sensitive content for compliance reporting.

Additional Endpoint Capabilities

The following features are highly desirable when deploying DLP at the endpoint:

- Endpoint agents and rules should be centrally managed by the same DLP management server that controls data in motion and data at rest (network and discovery).
- Policy creation and management should be fully integrated with other DLP policies in a single interface.
- Incidents should be reported to, and managed by, a central management server.
- Endpoint agent should use the same content analysis techniques and rules as the network servers/appliances.
- Rules (policies) should adjust based on where the endpoint is located (on or off the network). When the endpoint is on a managed network with gateway DLP, redundant local rules should be skipped to improve performance.
- Agent deployment should integrate with existing enterprise software deployment tools.
- Policy updates should offer options for secure management via the DLP management server, or existing enterprise software update tools.

Endpoint Limitations

Realistically, the performance and storage limitations of the endpoint will restrict the types of content analysis supported and the number and type of policies that are enforced locally. For some enterprises this might not matter, depending on the kinds of policies to be enforced, but in many cases endpoints impose significant constraints on data in use policies.

Central Administration, Policy Management, and Workflow

As we've discussed throughout this report, all current DLP solutions include a central management server for administering enforcement and detection points, creating and administering policies, incident workflow, and reporting. These features are frequently the most influential in the selection process. There are a lot of differences between the various products on the market; rather than trying to cover every possible feature, we'll focus on the baseline of functions that are most important.

User Interface

Unlike other security tools, DLP tools are often used by non-technical staff ranging from HR, to executive management, to corporate legal and business unit heads. As such the user interface needs to account for this mix of technical and non-technical staff and must be easily customizable to meet the needs of any particular user group. Due to the complexity and volume of information a DLP solution may deal with, the user interface can make or break a DLP product. For example, simply highlighting the portions of an email in violation of a policy when displaying the incident can shave minutes off handling time and avoid misanalyses. A DLP user interface should include the following elements:

- *Dashboard*: A good dashboard will have user-selectable elements and defaults for technical and non-technical users. Individual elements can be may only be available to authorized users or groups, typically groups stored in enterprise directories. The dashboard should focus on the information valuable to that user, and not be just a generic system-wide view. Obvious elements include number and distribution of violations based on severity and channel and other top-level information, to summarize the overall risk to the enterprise.
- *Incident Management Queue*: The incident management queue is the single most important component of the user interface. This is the screen incident handlers use to monitor and manage policy violations. The queue should be concise, customizable, and easy to read at a glance. Due to the importance of this feature we detail recommended functionality later in this report.
- *Single Incident Display*: When a handler digs into a single incident, the display should cleanly and concisely summarize the reason for the violation, the user involved, the criticality, the severity (criticality is based on which policy is violated, severity upon how much data is involved), related incidents, and all other information needed to make an intelligent decision on incident disposition.
- *System Administration*: Standard system status and administration interface, including user and group administration.
- *Hierarchical Administration*: Status and administration for remote components of the DLP solution, such as enforcement points, remote offices, and endpoints, including comparisons of which rules are active where.
- *Reporting*: A mix of customizable pre-built reports and tools to facilitate *ad hoc* reporting.

- *Policy Creation and Management:* After the incident queue, this is the most important element of the central management server. It includes the creation and management of policies. Because it's so important, we'll cover it in more detail later.

A DLP interface should be clean and easy to navigate. That may sound basic, but we're all far too familiar with poorly designed security tools that rely on the technical skills of the administrator to get around. Since DLP is used outside of security, possibly even outside of IT, the user interface needs to work for a wide range of users.

Hierarchical Management, Directory Integration, and Role-Based Administration

Hierarchical Management

DLP policies and enforcement often need to be tailored to the needs of individual business units or geographic locations. Hierarchical management allows you to establish multiple policy servers throughout the organization, with a hierarchy of administration and policies. For example, a geographic region might have its own policy server slaved to a central policy server. That region can create its own specific policies, ignore central policies with permission, and handle local incidents. Violations would be aggregated on the central server while certain policies are always enforced centrally. The DLP tool would support the creation of global and local policies, assign policies for local or global enforcement, and manage workflow and reporting across locations.

Directory Integration

DLP solutions also integrate with enterprise directories (typically Microsoft Active Directory) so violations can be tied to users, not IP addresses. This is complicated because they must deal with a mix of managed and unmanaged (guest/temporary) employees without assigned addresses. The integration should tie DHCP leases to users based on their network login, and update to avoid accidentally tying a policy violation to an innocent user. For example, an early version of a current product associated a user to an IP address until the address was reassigned to another user. One reference almost fired an employee because a contractor (not in Active Directory) was the next person to use that IP and committed a policy violation. The tool linked the violation to the innocent employee. Directory integration also streamlines incident management by eliminating the need to reference external data sources for user identities and organization structure.

Role-Based Administration

The system should also allow internal role-based administration for both internal administrative tasks and monitoring & enforcement. Internally, users can be assigned to administrative and policy groups for separation of duties. For example, someone might be given the role of enforcing any policy assigned to the accounting group, without access to administer the system, create policies, see violations for any other group, or alter policies. Since your Active Directory might not reflect the user categories needed for monitoring and enforcement, the DLP system should provide flexible support for monitoring and enforcement based on DLP-product specific groups and roles.

Policy Creation and Management

Policy creation and management is a critical function at the heart of DLP; it's also (potentially) the most difficult part of managing DLP. The policy creation interface should be accessible to both technical and non-technical users, although creation of heavily customized policies will nearly always require technical skills.

For policy creation, the system should let you identify the kind of data to protect, a source for the data if appropriate, destinations, which channels to monitor and protect, what actions to take for violations, which users to apply the policy to, and what handlers and administrators can access the policy and violations. Since not all policies are created equal, each should also be assigned a sensitivity, and severity thresholds, based upon volume of violations. Policies should be usable as templates for new policies, and if the system includes categories (which you really want) the policies associated with a given category should also be editable and available as templates for custom policies. Policy wizards are also a useful feature, especially for policies like protecting a single document.

Most users tend to prefer user interfaces that use clear, graphical layouts for policies, preferably with an easy-to-read grid of channels monitored and disposition for violations on that channel. The more complex a policy, the easier it is to create internal discrepancies or accidentally assign the wrong disposition to the wrong channel or violation.

Essentially every policy will need some level of tuning, and a good tool will allow you to create a policy in test mode that shows how it would react in production, without filling incident handlers' queues or taking any enforcement action. Some tools can test draft policies against previously recorded traffic.

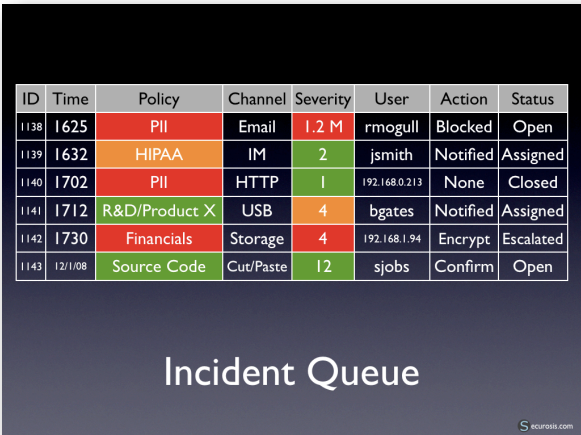
Policies include extremely sensitive information, so they should be hashed, encrypted, or otherwise protected within the system. Some business units may have extremely sensitive policies which will need to be protected against system administrators without explicit permission to see a particular policy. All policy violations should also be protected.

Incident Workflow and Case Management

Incident workflow will be the most heavily used part of the DLP system. This is where violations are reported, incidents are managed, and investigations are performed.

The first stop is the incident handling queue, which is a summary of all incidents either assigned to that handler, or unassigned but within the enforcement domain of the handler. Incident status should be clearly indicated with color-coded sensitivity (based on the policy violated) and severity (based on the volume of the transgression, or some other factor defined in the policy). Each incident should appear on a single line, and be sortable or filterable on any field.

Channel, policy violated, user, incident status (open, closed, assigned, unassigned, investigation) and handler should also be indicated and easily changed for instant disposition. By default, closed incidents shouldn't clutter the interface — basically treating it like an email Inbox. Each user should be able to customize anything to better suit his or her work style. Incidents with either multiple policy violations, or multiple violations of a single policy, should only appear once in the incident queue. An email with 10 attachments it shouldn't show up as 10 different incidents, unless each attachment violated a different policy.



ID	Time	Policy	Channel	Severity	User	Action	Status
1138	1625	PII	Email	1.2 M	rmogull	Blocked	Open
1139	1632	HIPAA	IM	2	jsmith	Notified	Assigned
1140	1702	PII	HTTP	1	192.168.0.213	None	Closed
1141	1712	R&D/Product X	USB	4	bgates	Notified	Assigned
1142	1730	Financials	Storage	4	192.168.1.94	Encrypt	Escalated
1143	12/1/08	Source Code	Cut/Paste	12	sjobs	Confirm	Open

Incident Queue

Securosis.com

When a single incident is opened, it should list all the incident details, including (unless otherwise restricted) highlighting what data in the document or traffic violated which policy. A valuable feature is a summary of other recent violations by that user, and other violations with that data (which could indicate a larger event). The tool should allow the handler to make comments, assign additional handlers, notify management, and upload any supporting documentation.

More advanced tools include case management for detailed tracking of an incident and any supporting documentation, including time-stamps and hashes of data. This is valuable in cases where legal action is taken, and evidence in the case management system should be managed to increase its suitability for admission in court.

System Administration, Reporting, and Other Features

As with any security tool, a DLP solution should include all the basic system administration features, including:

- Backup and restore — both full system and system configuration only for platform migrations.
- Import/Export — for policies and violations. There should be provision for extracting closed violations to free up space.
- Load balancing/clustering
- Performance monitoring and tuning
- Database management

Tools tend to mix these functions between the tool itself and the underlying platform. Some organizations will prefer to completely manage the tool internally without requiring the administrator to learn or manage the platform. As much as possible, you should look for a DLP tool that lets you manage everything through the included interface.

Reporting varies widely across solutions; some use internal reporting interfaces while others rely on third party tools like Crystal Reports. All tools ship with some default reports, but not all tools allow you to create your own reports. You should look for a mix of technical and non-technical reports, and if compliance is an issue consider tools that bundle compliance reports.

When you use data at rest or data in use (endpoint) features, you'll need a management interface that allows you to manage policies over servers, storage, and endpoints. The tool should support device grouping, rolling signature updates, and other features needed to manage large numbers of devices.

Beyond these basic features, products start to differentiate themselves with other advances to help meet particular enterprise needs, including:

- Third party integration, from web gateways to forensics tools.
- Language support, including double-byte character sets for Asia.
- Anonymization of policy violations to support international workplace privacy requirements.
- Full capture for recording all traffic, not just policy violations.

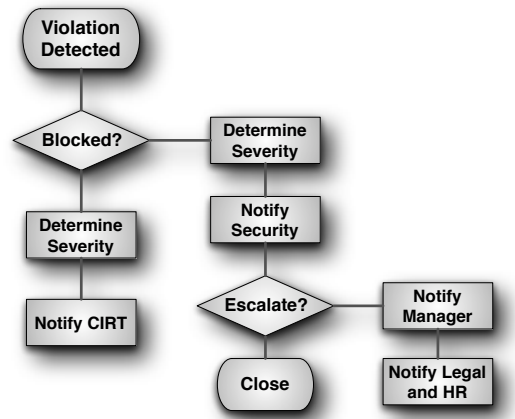
The DLP Selection Process

Define Needs and Prepare Your Organization

Before you start looking at any tools, you need to understand why you might need DLP, how you plan on using it, and the business processes around creating policies and managing incidents.

1. Identify business units that need to be involved and create a selection committee: We tend to include two kinds of business units in the DLP selection process — content owners with sensitive data to protect, and content protectors with the responsibility for enforcing controls over the data. Content owners include those business units that hold and use the data. Content protectors tend to include departments like Human Resources, IT Security, corporate Legal, Compliance, and Risk Management. Once you identify the major stakeholders, you'll want to bring them together for the next few steps.
2. Define what you want to protect: Start by listing out the kinds of data, as specifically as possible, that you plan on using DLP to protect. We typically break content out into three categories — personally identifiable information (PII, including healthcare, financial, and other data), corporate financial data, and intellectual property. The first two tend to be more structured and will drive you towards certain solutions, while IP tends to be less structured, with different content analysis requirements. Even if you want to protect all kinds of content, use this process to specify and prioritize, preferably “on paper”.
3. Decide how you want to protect it and set expectations: In this step you will answer two key questions. First, in what channels/phases do you want to protect the data? This is where you decide if you just want basic email monitoring, or if you want comprehensive data in motion, data at rest, and data in use protection. You should be extremely specific, listing out major network channels, data stores, and endpoint requirements. The second question is what kind of enforcement do you plan on implementing? Monitoring and alerting only? Email filtering? Automatic encryption? You'll get a little more specific in the formal requirements later, but you should have a good idea of your expectations at this point. Also, don't forget that needs change over time, so we recommend you break requirements into short term (within 6 months of deployment), mid-term (12-18 months after deployment), and long-term (up to 3 years after deployment).

4. Outline process workflow: One of the biggest stumbling blocks for DLP deployments is failure to prepare the enterprise. In this stage you define your expected workflows for creating new protection policies and handling incidents involving insiders and external attackers. Which business units are allowed to request protection of data? Who is responsible for building the policies? When a policy is violated, what's the workflow to remediate it? When is HR notified? Legal? Who handles day-to-day policy violations? Is it a technical security role, or non-technical, such as a compliance officer? The answers to these kinds of questions will guide you towards different solutions to meet your workflow needs.



By the completion of this phase you should have defined key stakeholders, convened a selection team, prioritized the data you want to protect, determined where you want to protect it, and roughed out workflow requirements for building policies and remediating incidents.

Formalize Requirements

This phase can be performed by a smaller team working under the mandate of the selection committee. Here, the generic needs determined in phase 1 are translated into specific technical features, while any additional requirements are considered. This is the time to come up with any criteria for directory integration, gateway integration, data storage, hierarchical deployments, endpoint integration, and so on. You can always refine these requirements after you proceed to the selection process and get a better feel for how the products work.

At the conclusion of this stage you develop a formal RFI (Request For Information) to release to vendors, and a rough RFP (Request For Proposals) that you'll clean up and formally issue in the evaluation phase.

Evaluate Products

As with any products, it's sometimes difficult to cut through marketing materials and figure out if a product really meets your needs. The following steps should minimize your risk and help you feel confident in your final decision:

1. *Issue the RFI:* Larger organizations should issue an RFI through established channels and contact a few leading DLP vendors directly. If you're a smaller organization, start by sending your RFI to a trusted VAR and email a few of the DLP vendors which seem appropriate for your organization.
2. *Perform a paper evaluation:* Before bringing anyone in, match any materials from the vendor or other sources to your RFI and draft RFP. Your goal is to build a short list of 3 products which match your needs. You should also use outside research sources and product comparisons.
3. *Bring in 3 vendors for an on-site presentation and risk assessment:* Nearly every DLP vendor will be happy to come in and install their product on your network in monitoring mode for a few days, with a suite of basic rules. You'll want to try and overlap the products as much as possible to directly compare results based on the same traffic over the same time period. This is also your first chance to meet directly with the vendors (or your VAR) and get more specific answers to any questions. Some vendors may (legitimately) desire a formal RFP before dedicating resources for any on-site demonstrations, especially for smaller organizations.
4. *Finalize your RFP and issue it to your short list of vendors:* At this point you should completely understand your specific requirements and issue a formal RFP.

5. *Assess RFP responses and begin product testing:* Review the RFP results and drop anyone who doesn't meet any of your minimal requirements (such as directory integration), as opposed to "nice to have" features. Then bring in any remaining products for in-house testing. To properly test products, place them on your network in passive monitor mode and load up some sample rule-sets that represent the kinds of rules you'd deploy in production. This lets you compare products side by side, running equivalent rules, on the same traffic. You'll also want to test any other specific features that are high on your priority list.
6. *Select, negotiate, and buy:* Finish testing, take the results to the full selection committee, and begin negotiating with your top choice.

Internal Testing

In-house testing is the last chance to find problems in your selection process. Make sure you test the products as thoroughly as possible. A few key aspects to test, if you can, are:

- Policy creation and content analysis. Violate policies and try to evade or overwhelm the tool to learn where its limits are.
- Email integration.
- Incident workflow — Review the working interface with those employees who will be responsible for enforcement.
- Directory integration.
- Storage integration on major platforms to test performance and compatibility for data at rest protection.
- Endpoint functionality on your standard image.
- Network performance — not just bandwidth, but any requirements to integrate the product on your network and tune it. Do you need to pre-filter traffic? Do you need to specify port and protocol combinations?
- Network gateway integration.
- Enforcement actions.

Conclusion

Navigating the Maze

Data Loss Prevention is a confusing market, but by understanding the capabilities of DLP tools and using a structured selection process you can choose an appropriate tool for your requirements.

I've worked with a hundred or more organizations evaluating DLP since the inception of the market. Not all of them bought a product, and not all of them implemented one, but those which did generally found the implementation easier than many other security products. From a technical standpoint that is; the biggest obstacles to a successful DLP deployment tend to be inappropriate expectations and failing to prepare for the business process and workflow of DLP.

Set your expectations properly. DLP is a very effective tool for preventing accidental disclosures and ending bad business processes around the use of sensitive data. While it offers some protection against malicious attacks, the market is still a couple years away from stopping knowledgeable bad guys.

Have a clear understanding of which business units will be involved and how you plan to deal with violations before you begin the selection process. After deployment is a bad time to realize that the wrong people see policy violations, or your new purchase isn't capable of protecting the sensitive data of a business unit not included in the selection process.

DLP products may be adolescent, but they provide very high value for organizations that plan properly and understand how to take full advantage of them. Focus on those features that are most important to you as an organization, paying particular attention to policy creation and workflow, and work with key business units early in the process.

Most organizations that deploy DLP see significant risk reductions with few false positives and business interruptions.

About the Author

Rich Mogull has over 17 years experience in information security, physical security, and risk management. Prior to founding Securosis, Rich spent 7 years as a leading security analyst with Gartner, where he advised thousands of clients, authored dozens of reports, and was consistently rated one of Gartner's top international speakers. He is one of the world's premier authorities on data security technologies, and has covered issues ranging from vulnerabilities and threats, to risk management frameworks, to major application security.

About Securosis

Securosis, L.L.C. is the independent security consulting practice of Rich Mogull.

Securosis provides security consulting services in a variety of areas, including:

- Security Management and Strategy
- Technology Evaluations (for end users and investors)
- Product Selection Assistance
- Security Market Strategies
- Risk Management and Risk Assessment
- Data Security Architecture
- Security Awareness and Education

Securosis partners with security testing labs to provide unique product evaluations that combine in-depth technical analysis with high-level product, architecture, and market analysis.

About the SANS Institute

SANS is the most trusted and by far the largest source for [information security](#) training and certification in the world. It also develops, maintains, and makes available at no cost, the largest collection of research documents about various aspects of information security, and it operates the Internet's early warning system -- the [Internet Storm Center](#).

Many of the valuable SANS resources are free to all who ask. They include the very popular Internet Storm Center, a weekly news digest ([NewsBites](#)), a weekly vulnerability digest ([@RISK](#)), flash security alerts, and more than 1,200 award-winning original [research papers](#).