Understanding and Selecting a Data Loss Prevention Solution

Version 3.0
Released: Oct 31, 2017

Securosis, L.L.C. 515 E. Carefree Highway Suite #766 Phoenix, AZ 85085 **T** 602-412-3051
info@securosis.com www.securosis.com

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 and professionally edited.

Special thanks to Chris Pepper for editing and content support.

Understanding and Selecting DLP

Table of Contents

Introduction to DLP

A (Still) Confusing Market

Data Loss Prevention has matured considerably over the past decade, but selecting DLP technology can still be confusing. Some aspects of DLP have appeared in a variety of other product categories as value-add features, blurring the lines between purpose-built DLP solutions and traditional security controls, including next generation firewalls and email security gateways. Meanwhile purpose-built DLP tools continued to evolve — expanding coverage, features, and capabilities to address advanced and innovative means of exfiltrating data.

Even today it can still be difficult to understand the value of the various tools, and which products best suit which environments — further complicated by the wide variety of deployment models. You can go with a full-suite solution that covers your network, storage infrastructure, and endpoints; or focus on a single 'channel'. You might already have content analysis and policy enforcement directly embedded into your firewall, web gateway, email security service, CASB, or other tools.

So the question is no longer only "Do I need DLP and which product should I buy?" but "What kind of DLP will work best for my needs, and how can I figure that out?" This paper provides the background on DLP to help you understand the technology, know what to look for in a product (or service), and find the best match for your organization.

Defining DLP

Even a decade on, there is still little consensus on what actually compromises a DLP solution. Some people consider encryption or USB port control to be DLP, while others limit the term to complete product suites focused on analyzing and enforcing content usage policies. 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 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 an enterprise. Few enterprises classify data beyond public vs. everything else. DLP helps organizations better understand their data, and improves their ability to classify and manage content.

Full-suite solutions provide complete coverage across your network, storage repositories, and endpoints, even if you aren't using their full capabilities. There are three other approaches:

- *Partial-suite* DLP solutions are dedicated DLP tools that cover two potential channels (such as network and storage) with full workflow (such as case management) and content analysis capabilities.
- *Single-channel* DLP solutions cover only one channel, but still include full DLP workflow and content analysis.
- *DLP features* are now included in a variety of products, offering a subset of coverage and content analysis capabilities, and typically lack dedicated DLP workflow. For example we have seen next generation firewalls and web gateways with basic pattern-matching capabilities, vulnerability assessment scanners which look for particular data types such as credit card numbers, and limited content analysis in an email security gateway.

DLP Features vs. DLP Solutions

When evaluating options it can be difficult to characterize the real differences between DLP features and dedicated DLP solutions, and the value of each. The key differences are:

- *A DLP product or solution* includes centralized management, policy creation, and enforcement workflow, dedicated to the monitoring and protection of content and data with full content awareness of the types of data that you need to protect. The user interface and functionality are dedicated to solving the business and technical problems of protecting content.
- *DLP features* include some of the detection and enforcement capabilities of DLP products, but are not embodied in products dedicated to protecting content and data.

The latter is sometimes called "DLP Light" to reflect its less robust nature, and is becoming extremely common across a variety of other security tools.

This distinction is important because DLP products solve a specific business problem which 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, and therefore the primary user of the DLP solution. Even human resources is often involved in 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 security, which also may argue for a dedicated solution. Because DLP is dedicated to a clear business problem (protect my content) which is differentiated from other security problems (protect my PC or protect my network), if your primary goal is data protection you should focus on DLP solutions rather than features in products with different focus.

This doesn't mean DLP Light cannot be the right solution for your requirements, especially in smaller organizations. It tends to cost less and is often as easy to deploy as turning a switch on a management console. If you only need basic credit card protection and can accept coverage, analysis, and workflow limitations, the DLP features of another product might meet your needs. Or you might start with a DLP feature of a product already in your environment to dip your toes in the water and get a better sense of how big a problem you have, before migrating to a full dedicated solution. We will provide more direct guidance on how to choose between the two below under *DLP Selection Process*, and then describe common implementations under *Technical Architecture*.

Content Awareness

Content vs. Context

Before delving into the particulars of different content analysis techniques, we need to distinguish content from context. One of the defining characteristics of DLP solutions is *content awareness*. This is the ability to deeply analyze content using a variety of techniques, and is very different from analyzing context. It is easiest to think of content as a letter, and context as the envelope and environment around it. Context includes source, destination, size, recipients, sender, header information, metadata, time, format, and anything else short of the content of the letter itself. Context is essential, and any DLP solution should include contextual analysis.

A more advanced version of contextual analysis is *business context analysis,* which involves deeper analysis of content, its environment at the time of analysis, and the use of the content at that time. For example while an envelope might tell you sender and destination, business context can tell you which business unit the current holder of the envelope belongs to, their virtual location, what application they are reading it with, and so on.

Content awareness requires peering inside containers to analyze the content itself. It is the next level of intelligence beyond context. If I want to protect a piece of sensitive data, I want to protect it everywhere — not just in obviously sensitive containers. I need to protecting the data, not the envelope, so it's very useful to open the letter, read it, and *then* decide how to handle it. This is more difficult and time-consuming than basic contextual analysis, and the defining characteristic of DLP solutions.

Contextual Analysis

Early contextual analysis used to be pretty simple — often little more than scanning email headers or file metadata. Since then it has evolved considerably to evaluate factors such as:

- File ownership and permissions.
- Use of encrypted file formats or network protocols.
- User role and business unit (through directory integration).
- Specific web services — such as known webmail providers and social networking sites.
- Web addresses (not just the session content).
- USB device information, such as manufacturer or model number.
- The desktop application in use (*e.g.,* knowing something was copied from an Office document and then pasted into an encryption tool).

Contextual analysis often provides *business context* for subsequent content analysis. This is one of the major benefits of DLP: rather than looking at packets or files in a vacuum, you can build policies which take into account everything from the employee's job or role, to the application in use.

Content Analysis

The first step in content analysis is capturing the envelope and opening it. The DLP engine then needs to parse the context and dig into it. For a plain text email this is easy, but looking inside binary files gets a bit more complicated. All DLP solutions solve this using *file cracking*. File cracking is the technology used to read and understand a file, even if its content is buried under multiple layers. For example it is common for a cracker to read an Excel spreadsheet embedded in a zipped Word file. 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 more complex, such as a PDF embedded in a CAD file. Many products on the market today support around 300 file types, embedded content, multiple languages, double-byte character sets for Asian languages, and extracting plain text from unidentified file types. Some tools can analyze encrypted data if enterprise encryption recovery keys are used. And most can identify standard encryption and use that as a cue to block or quarantine content.

Content Analysis Techniques

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

1. **Rules-Based/Regular Expressions:** This is the most common analysis technique, available in both DLP products and DLP features inside other tools. It compares content using specific rules — such as 16-digit numbers which meet credit card checksum requirements, medical billing codes, and other textual analyses. Most DLP solutions enhance basic regular expressions with their own additional analyses (*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.

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

   Weaknesses: Prone to higher false positive rates. Offers very little protection for unstructured content such as sensitive intellectual property.

2. **Database Fingerprinting:** Also called Exact Data Matching, this technique takes either a database dump or live data from a database (via ODBC connection) and looks only for exact matches. 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, which triggers most US state breach disclosure laws. 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 few false positives (close to 0). Enables you to protect customer and other sensitive data while ignoring other similar data used by employees, such as their personal credit cards.

Weaknesses: Nightly dumps don't include 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 hashes of important files, and then monitor for any files matching any of those fingerprints. Some people consider this contextual analysis, because the file contents themselves are not analyzed.

   What it's best for: Media files and other binaries where textual analysis isn't necessarily possible, such as photos, audio, movies, and certain proprietary design and engineering files.

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

   Weaknesses: High false negatives because this approach is trivial to evade. Worthless for content that has been edited, such as standard office documents and edited media files. Adversaries can pack files to change their hashes with little effort.

4. **Partial Document Matching:** This technique looks for complete or partial matches to protected content. You could build a policy to protect a sensitive document, and your DLP solution can look for either its complete text or excerpts as small as a few sentences. For example you could load up a business plan for a new product, and a DLP solution could alert if any employee pastes a single paragraph into a chat window.

   What it's best for: Protecting sensitive documents or similar content, including source code. Unstructured content which is known to be sensitive.

   Strengths: Ability to protect unstructured data. Generally low false positives (some vendors 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 or verbiage in a protected document may trigger false positives. Must decide exactly which documents to protect. As discussed above, file matching is trivial to avoid.

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 protected content. This category includes a wide range of statistical techniques which vary greatly in implementation and effectiveness.

   What it's best for: Unstructured content where a deterministic technique such as 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 extreme 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 negatives. Requires a large corpus of source content to train the algorithms — 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 which resembles insider trading — using key phrases, word counts, and positions to find violations. Other examples include 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 but are similar to 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 which other techniques can't even attempt to monitor.

   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 both false positives and 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.

   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 data protection needs.

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

These 7 techniques serve as the basis for DLP products on the market. Not all products include all techniques, and you will find significant differences between implementations. Most products also support chaining techniques: building complex policies from combinations of content and contextual analyses. This basically uses a cocktail of analysis techniques to increase accuracy.

Technical Architecture

Protecting Data in Motion, at Rest, and in Use

The goal of DLP is to protect content throughout its lifecycle — on the network, in storage, and on endpoints. This includes three major aspects:

- **Data in Motion** protection is monitoring (and potentially filtering) traffic on the network (passively or inline via proxy) to identify content being sent across specific communications channels. This includes monitoring email and web traffic (including apps encapsulating their traffic on port 80) for snippets of sensitive content. These tools can often block several types of traffic based on central policies.
- **Data at Rest** is protected by scanning storage and other content repositories to identify where sensitive content is located. We often 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 an alert sent to the file owner and the SOC.
- **Data in Use** is addressed by endpoint solutions that monitor data as users interact 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 USB drives entirely), stop transfer of a file to a webmail system, or prevent a file from being printed. Data in use tools can also detect things like copy and paste, as well as use of sensitive data in unapproved applications (such as someone attempting to encrypt data to sneak it past sensors).

As we hinted above, these translate to three major architectural components: *network monitoring/filtering, storage scanning,* and *endpoint agents.* Even DLP features of other security products or DLP services fit one or more of these major architectures.

The rest of this section will focus on dedicated DLP solutions, but many of these details also apply to DLP features within other products. We will also talk more about common DLP Light architectures.

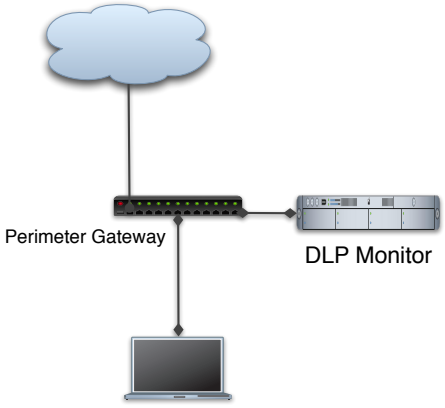
Networks

Many organizations first enter the world of DLP with network products which provide broad protection for managed and unmanaged systems. It's typically easier to start a deployment with network products to gain broad coverage quickly because that doesn't require implementing anything on each device.

Network Monitor

At the heart of most DLP solutions lies a passive network monitor. The network monitoring component is typically deployed at or near a gateway on a SPAN or mirror port or 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. Many clients claim they need performance to match their full peak bandwidth, but that level of throughput is unnecessary except in very unusual circumstances. Maximum possible bandwidth is generally substantially higher than realistic utilization, so be careful not to waste a large amount of money over provisioning DLP. And remember that DLP is for monitoring employees, not indiscriminately scanning *all* network traffic.



Passive Monitoring Architecture

You might have to make a trade-off between filtering the data that goes into the DLP (and missing non-standard traffic) or buying more boxes and load balancing. Some products lock monitoring into predefined port and protocol combinations, rather than using service/channel identification based on packet content. Make sure the network DLP is looking at the entire network stack (all ports/protocols), otherwise you might miss non-standard communications such as connecting over an unusual port.

Keep in mind, especially when testing, that *performance is often tied to the number and scope of DLP policies you deploy*. If you perform large amounts of partial document matching or database fingerprinting you might find performance slowing and need to move toward load balancing or separating traffic streams. You might offload email to a dedicated email monitoring system because email uses a different architecture than web and most other traffic.

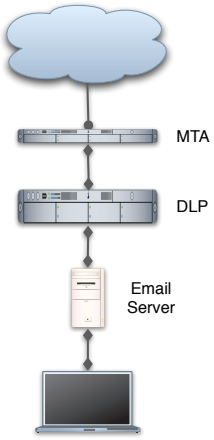
Finally, there are some organizations with outsize average monitoring requirements in terms of both bandwidth and port/protocol coverage, due to either size or the nature of threat they face — advanced attackers are more likely to exfiltrate data over unusual ports & protocols. If you fall into this category make sure to include it in your DLP requirements and test it during the selection process. You should also coordinate your DLP program with any other egress filtering projects, which may be able to reduce your DLP load.

Email Integration

The next major component is email integration. The latency tolerant store and forward architecture of email enables several additional capabilities including quarantine, encryption integration, and filtering — without the tight throughput constraints required to avoid blocking synchronous traffic. Most products embed an MTA (Mail Transport Agent), allowing you to simply add it as another hop in the email chain. Quite a few also integrate directly with some of the major existing MTAs and email security services for better performance. To monitor internal mail you need direct email system integration.

Filtering/Blocking and Proxy Integration

Nearly anyone deploying a DLP solution will eventually want to start blocking traffic. You can watch all your juicy sensitive data flowing out to the nether regions of the Internet for only so long before you need to start taking action. But blocking isn't the easiest thing in the world — you need to allow all good traffic, block only bad traffic, and make the decision using real-time content analysis. Oh, and you can't afford mistakes.

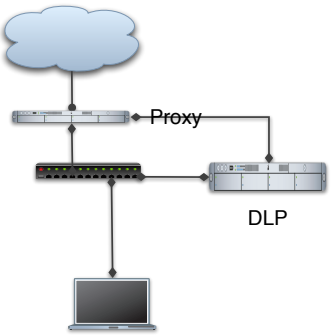


Email Architecture

Email, as just mentioned, is fairly straightforward to filter. It's not quite real-time and proxied by nature. Adding one more analysis hop is a manageable problem in even the most complex environments. Outside of email, however, most communications traffic is sensitive to latency because everything runs in real time. So to filter we either need to proxy traffic or kill bad sessions from the outside.

Proxy

In simplified terms, a proxy is a protocol/application specific application which queues up traffic before passing it on, allowing for deeper analysis. Few DLP solutions include their own proxies — they tend to integrate with existing gateway/proxy vendors because most customers prefer integration with tools they have already deployed. Integration for web gateways is typically through the iCAP protocol — this enables the proxy to grab the traffic, send it to the DLP product for analysis, and terminate communications if there's a violation. This means (assuming you already use an iCAP compatible gateway) you don't need to add another piece of hardware in front of your network traffic, and DLP vendors can avoid building dedicated network hardware for inline analysis.



Proxy/Gateway

Be sure to look for a reverse SSL proxy to allow inspection of SSL connections, although TLS 1.3 does impact how encrypted sessions can be decrypted and will push more reliance on endpoint DLP analysis. You will need to make changes on your endpoints to deal with all the certificate alerts, but can now peer into encrypted traffic.

To reiterate, we *highly* recommend that if you monitor web traffic, you deploy and integrate your DLP with a reverse SSL proxy so you can view encrypted traffic, give that a majority of network traffic will be encrypted within the next couple of years.

Service or On-Premise

With the advent of SECaaS (Security as a Service) solutions, it has become common to use a service to proxy outbound web and email traffic. The question regarding SECaaS solutions is how deep and granular solution can go. Some providers run a dedicated DLP environment for you in their cloud. Others offer content analysis capabilities in web and email proxy services. So the same kinds of rules apply when selecting a service as when selecting an on-premise solution.

TCP Poisoning

The last method we'll discuss is TCP poisoning. You monitor traffic, and when you see something bad you inject a TCP reset packet to kill that connection. This works on every TCP protocol but isn't very efficient. For one thing some protocols keep trying to get traffic through. If you TCP poison a single email message the server will keep trying to send it for several days, as often as every 15 minutes. The other problem is having to analyze in real time since you don't queue traffic at all, by the time you notice something bad it might be too late. It's a good stopgap for broad protocol coverage but for filtering you should rely on proxies as much as possible.

Internal Networks

Although technically capable of monitoring internal networks, DLP is rarely used on internal traffic other than email. Perimeter gateways provide convenient choke points — without them internal monitoring is a daunting prospect due to cost, performance, policy management and false positives.

Distributed and Hierarchical Deployments

All medium to large enterprises, and many smaller organizations, have multiple locations and Internet egress points. 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, investigation, and archiving. Remote offices are usually easy to support because you can just push policies down and reporting back up, so make sure your product offers this capability.

More advanced products support hierarchical deployments for organizations which want to manage DLP differently between multiple geographic locations or by business unit. International companies often need this to meet privacy protection requirements which vary by country. Hierarchical management supports coordinated local policies and enforcement in different regions, running on their own local servers, communicating back to a central management server.

Managed service deployments may involve running the management server/aggregation point outside your network. There is nothing fundamentally wrong with this, but you should be diligent to ensure there is sufficient data isolation and your DLP console is sufficiently protected. With so much sensitive information moving around it is important that each tier in the hierarchy is well secured and all communications between DLP nodes is encrypted, regardless of where they reside.

Storage

While catching leaks on the network before data is lost is very powerful, it's only one aspect of the challenge of data protection. The first step is to figure out where all that sensitive data is stored in the first place. We call this *content discovery.* Enterprise search or electronic discovery tools might be able to help with this, but they aren't tuned for this specific problem. Enterprise data classification tools can also help, but don't seem to work well for finding specific policy violations. So we see many clients leveraging the content discovery features of DLP products.

Content discovery consists of three components, based on where information is stored:

1. **Storage:** scanning mass storage including file servers, SAN, and NAS.
2. **Applications/Servers:** application-specific scanning of stored data on email servers, document management systems, and databases.
3. **Endpoints:** scanning workstations, laptops, and mobile devices.

We will focus mostly on storage and application/server scanning in this section; we will cover endpoints in detail later.

Content Discovery Techniques

There are four basic techniques for content discovery:

1. **Remote Scanning:** Either the central policy server or a dedicated scanning server accesses storage repositories via network shares or other administrative access. Files are scanned for content violations. Connections often use administrative credentials. Any content transferred should be encrypted, but this may require reconfiguration of the storage repository and isn't always feasible. Most tools allow bandwidth throttling to limit network impact, and scanning servers are often placed close to storage to increase speed and limit network impact. This technology supports scanning nearly any storage repository, but performance is limited by the network, even with optimization.
2. **Agent-Based Scanning:** An agent is installed on the system (endpoint/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.
3. **Memory-Resident Agent Scanning:** Rather than deploying a persistent agent, a memory-resident agent is installed which performs a scan and then exits without leaving anything running or stored on the local system. This offers the performance and granularity of agent-based scanning without an agent running all the time.
4. **Application Integration:** Direct integration (often using an agent) with document management, content management, or other storage repositories. This integration not only supports visibility into management content, but allows the discovery tool to understand local context and metadata, and possibly enforce actions within the system.

Enterprises typically deploy a mix of these technologies depending on policy and infrastructure requirements. We currently see deployments guided by technology limitations of each approach:

- Remote scanning can significantly increase network traffic and has performance limitations based on network bandwidth and target and scanner network performance. These physical realities present practical limitations — which may rule out some options in large storage environments. The advantage is that you only need access to a file share.
- Agents, ephemeral or permanent, are limited by processing power and memory on the target system, which may translate into restrictions on the number of policies that can be enforced, and the types of content analysis which can be used without adversely impacting performance.
- Agents don't support all platforms.

Storage Enforcement

Once a policy violation is discovered, a 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 or popup on their device 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 for 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 simply 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. For example we often see deployments of DLP 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.

Integration and Additional Features

The most common integration we see is with document management systems (DMS). When integrated with a DMS, DLP offers more information than typically provided by a file share, sometimes even including file usage. DMS metadata and privileges are usually more robust than those on generic file shares, providing greater context for policies.

On the database side most products include the ability to look for sensitive information over an ODBC connection or direct support for the database. Rather than looking at *all* data in the database, the tool scans the first *n* rows of a table and column headers to see if any sensitive data might be present. This is especially useful for evaluating the many *ad hoc* databases commonly found within business units.

Endpoints

DLP usually starts on the network because that's the most cost-effective way to get the broadest coverage with the least organizational friction. Network DLP is not intrusive (unless you have to crack SSL) and offers visibility into any system on the network, managed or unmanaged, server or workstation. Filtering is more difficult but still relatively straightforward (especially for email) and covers all systems connected to the network. But clearly 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 also the endpoints where data is used.

> *Note: Despite significant advancements in endpoint DLP, endpoint-only solutions are not recommended for most users. As we'll discuss, they normally require compromise on the numbers and types of policies that can be enforced, offer limited email integration, and offer no protection for unmanaged systems. Long-term you need both network and endpoint capabilities, and nearly every network DLP solution offers at least some endpoint protection.*

Adding an endpoint agent to a DLP solution not only gives you the ability to discover locally stored content, but also to potentially protect systems no longer on the network, or even protect data as being actively used. While extremely powerful, endpoint agents can be problematic. 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” which generates false positives every time an employee visits Amazon.com.

Key Capabilities

Agents include four generic layers/features:

1. **Content Discovery:** Scanning of stored content for policy violations.
2. **File System Protection:** Monitoring of and enforcement on file operations as they occur (as opposed to discovery, which is scanning of content already written to media). This is most often used to prevent content from being written to portable media (USB). It’s also where tools hook in for automatic encryption.
3. **Network Protection:** Monitoring and enforcement of network operations. An endpoint agent can provide protection similar to gateway DLP when an endpoint is off the corporate network. Most endpoints handle printing and faxing as network traffic, so this is where most print/fax protection can be enforced (the rest comes from special print/fax hooks).
4. **GUI/Kernel Protection:** A more generic category to cover data-in-use scenarios, such as Copy & Paste, application restrictions, and Print Screen.

Between these four categories we cover most of the day-to-day operations users perform on their devices which place content at risk. They address our primary drivers from the last section: protecting data from being copied to portable storage, protecting systems off the corporate network, and supporting discovery on endpoints. Most tools start with file and then networking features, before moving on to some of the more complex GUI/kernel functions.

Use Cases

Endpoint DLP needs to support these critical use cases:

- Ongoing scanning of local storage for sensitive data that shouldn’t ever appear on an endpoint, such as credit cards or customer Personally Identifiable Information (PII).
- Enforcing network rules off the managed network, including modified rules on more hostile networks.
- Restricting sensitive content from portable storage, including USB drives, CD/DVD drives, home storage, and devices such as 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, but not tools downloaded online that don’t support enterprise data recovery.
- Auditing use of sensitive content for compliance reporting.

Agent Content Awareness

Even if you have an endpoint with a huge processor and gigabytes of RAM, it would be wasteful to devote all that horsepower to enforcing DLP — especially if it interferes with the primary purpose of the system.

Content analysis may be resource intensive, depending on the types of policies to enforce. Additionally, different agents have different enforcement capabilities, which do not always match up to their gateway counterparts. At minimum most endpoint tools support rules/regular expressions, some degree of partial document matching, and a whole lot of contextual analysis. Others support their entire repertoire of content analysis techniques, but you will likely have to tune policies to run on more constrained endpoints.

Some tools rely on the central management server for certain aspects of content analysis. Rather than performing all analysis locally they ship content back to the server and act on any results. This obviously isn't ideal because such policies cannot be enforced when the endpoint is off the enterprise network, and shipping the data around sucks up a bit of bandwidth. But this does enable enforcement of policies that are otherwise totally unrealistic on an endpoint, such as fingerprinting of a large enterprise database.

A way to address this issue is to implement adaptable policies based on endpoint location. For example, when you're on the enterprise network most policies are enforced at the gateway. Once you access the Internet outside the corporate walls, a different set of policies is enforced. You might use database fingerprinting of the customer database at the gateway when the laptop is in the office or on a VPN, but drop to a rule/regex for Social Security Numbers or account numbers for mobile workers. Sure, you'll get more false positives, but you're still able to protect your sensitive information within performance constraints.

Agent Management

Agent management consists of two main functions: deployment and maintenance. On the deployment side most tools today are designed to work with whatever endpoint management tools your organization already uses. As with other software tools you create a deployment package and then distribute it along with any other software updates. If you don't already have a software deployment tool you'll want to look for an endpoint DLP tool which includes basic deployment capabilities. Because all endpoint DLP tools include central policy management, deployment is fairly straightforward. There's little need to customize packages based on user, group, or other variables beyond the location of the central management server.

The rest of the agent's lifecycle, aside from major updates, is controlled through the central management server, which may reside on-premise or in the cloud. Agents should communicate regularly with the central server to receive policy updates and report incidents & activity. When the central management server is accessible, this should happen near real-time. When the endpoint is not connected to a network the DLP tool will store violations locally in a secure repository that's encrypted and inaccessible to the user. The tool will connect to the management server next time it's accessible, receiving policy updates and reporting activity. The management server should produce aging reports to help you identify endpoints which are out of date and need to be refreshed. Under some circumstances the endpoint may be able to communicate remote violations through encrypted email or another secure mechanism from outside the corporate firewall.

Aside from content policy updates and activity reporting, a few other features require central management. For content discovery you'll need to control scanning schedule/frequency as well as bandwidth and performance (*e.g.*, capping CPU usage). For real-time monitoring and enforcement you'll also want performance controls, including limits on how much space is used to store policies and the local cache of incident information.

Once you establish your base configuration you shouldn't need to perform much endpoint management directly. Things like enforcement actions are handled implicitly as part of policy, so integrated into the main DLP policy interface.

Enforcement Options

Because any given endpoint agent might be monitoring data in motion, at rest, and in use, there is a wide range of alerting and enforcement options. Rather than listing every possible option, many of which are also available in network and storage DLP, here are some endpoint-specific examples:

- Blocking an employee from transferring a file to portable storage.
- Allowing transfer to portable storage but requiring the user to submit a "business justification" in a pop-up window, which is sent with the incident information to the central DLP server.
- Creating hidden 'shadow' copies of any files transferred to portable storage, which are sent to the DLP server the next time the user is on the organization's network and later reviewed to determine whether investigation is warranted.
- Allowing Copy & Paste only between approved applications (application control). Attempting to Copy & Paste protected content into an unapproved application is blocked (*e.g.,* to prevent pasting into an encryption application).
- Only allowing printing to approved print servers.

DLP Features and Integration with Other Security Products

Up to now we have mostly focused on describing aspects of dedicated DLP solutions, but we see increasing interest in DLP Light tools for four main use cases:

- Organizations which turn on the DLP feature of an existing security product, like an endpoint suite or IPS, to generally assess their data security issues. Users typically activate a few general rules and use the results more to scope out their issues than to actively enforce policies.
- Organizations which only need basic protection on one or a few channels for limited data types, and want to bundle in DLP with existing tools if possible — often to reduce cost. The most common examples are email filtering, endpoint storage monitoring, or content-based USB alerting/blocking for credit card numbers or customer PII.
- Organizations which want to dip their toes into DLP with plans for later expansion. They usually turn on the DLP features of an existing security tool which is also integrated with a larger DLP solution. These are often provided by larger vendors which have acquired a DLP solution and integrated certain features into an existing product line.
- To address a very specific and narrow compliance deficiency that a DLP Light feature can resolve.

There are others but these are the cases we encounter most often. DLP Light tends to work best when protection scope and content analysis requirements are limited, and cost is a major concern.

Although there are a myriad of options, we see some consistencies between the various DLP Light offerings, as well as full DLP integration with other existing tools. Here we will highlight the most common features and architectures, including places where full DLP solutions can integrate with existing infrastructure.

Content Analysis and Workflow

Most DLP Light tools start with some form of rules and pattern matching — usually regular expressions, often with some additional contextual analysis. This base feature covers everything from keywords to credit card numbers. Because most customers don't want to build their own custom rules, the tools come with pre-built policies. The most common is still to find credit card data for PCI compliance. Next we tend to see PII detection, followed by healthcare/HIPAA data discovery; all these use cases target clear compliance requirements.

The longer a tool or feature has been on the market, the more categories it tends to support, but few DLP Light tools or features support the more advanced content analysis techniques described in this paper. This usually means more false positives than a dedicated solution, but for some of these data types such as credit card numbers, even a false positive is something you will usually want to take a look at.

DLP Light tools or features also tend to be more limited in terms of workflow. They rarely provide dedicated DLP workflow, and policy alerts are integrated into whatever existing console and workflow the tool uses for its primary function. This might not be an issue, but it's definitely important to consider before making a final decision.

Network Features and Integration

DLP features are increasingly integrated into existing network security tools, especially email security gateways. The most common examples are:

- **Email Security Gateways:** These were the first non-DLP tools to include content analysis, and tend to offer the broadest policy/category coverage. Email gateways are also one of the top integration points for full DLP solutions: all policies and workflow are managed on the DLP side, but analysis and enforcement are integrated with the gateway directly rather than requiring a separate mail hop.
- **Web Security Gateways:** Secure web gateways now directly enforce DLP policies on the content they proxy, such as preventing files with credit card numbers from being uploaded to webmail and social networking services. Web proxies are the second most common integration point for DLP solutions because, as we described under *Technical Architecture*, they proxy web (and FTP) traffic and make a perfect filtering and enforcement point. These are also the tools you use to reverse proxy SSL connections to monitor encrypted communications, a critical capability these tools require to block inbound malicious content. Web gateways also provide valuable context, by categorizing URLs and web services to support policies which take account the web destination into account — not just content and port/protocol.

- **Next Generation Firewalls:** NGFWs provide broad network security coverage, including application-aware firewall and threat detection/IPS capabilities, as well as web filtering. These are a natural location to add network DLP coverage, for the same reasons as Web Security Gateways.

Endpoint Features and Integration

DLP features have appeared in various endpoint tools aside from dedicated DLP products since practically before there was much of a DLP market. This continues to expand, especially as interest grows in controlling USB usage without onerous business impact.

- **USB/Portable Device Control:** A frequent inhibitor to deployment of portable storage management tools is their impact on standard business processes. There is always a subset of users who legitimately need access to portable storage for file exchange (*e.g.,* sales presentations), but organizations still want to audit or even block inappropriate transfers. Even basic content awareness can help provide protection with reduced business impact. Some tools include basic DLP capabilities, and we are seeing others evolve to offer somewhat extensive endpoint DLP coverage — with multiple detection techniques, multivariate policies, and even dedicated workflow. When evaluating this option, keep in mind that some tools position themselves as offering DLP capabilities but lack *any* content analysis — instead relying solely on metadata or other context.
- **Endpoint Protection Platforms:** For those of you who don't know, EPP is the term for comprehensive endpoint suites that include antivirus, host intrusion prevention, and possibly other capabilities ranging from remote access and Network Admission Control to application whitelisting. Many EPP vendors have acquired full or endpoint-only DLP products and integrated those capabilities with their EPP suites. Other EPP vendors have added basic DLP features — most often for monitoring local files or storage transfers for sensitive information. Options are available for basic endpoint DLP (usually using a few preset categories), all the way up to a DLP client integrated with a dedicated DLP.
- **Advanced Endpoint Protection/EDR:** Emerging endpoint security technologies focus on detecting and responding to advanced attacks, yet don't claim to replace existing EPP (yet). These companies are on a collision course with endpoint DLP (and *vice versa*), because the inspection required to detect advanced malware is not fundamentally different from detecting content leakage.

Overall, most people deploying DLP features on an endpoint (without a dedicated DLP solution) are focused on scanning the local drive and/or monitoring/filtering file transfers to portable storage. But as described earlier, you might see anything from network filtering to application control integrated into endpoint tools.

Storage Features and Integration

We don't see nearly as much DLP Light in storage as in networking and endpoints — in large part because there aren't as many clear security integration points. Fewer organizations have any sort of storage security monitoring, whereas nearly every organization performs network and endpoint monitoring of some sort. But while we see less DLP Light, as we already discussed, we see extensive integration on the DLP side for different types of storage repositories.

- **Database Activity Monitoring and Vulnerability Assessment:** DAM products, which now include Database Vulnerability Assessment tools, sometimes include content analysis capabilities. These are designed to find sensitive data in large databases, detect sensitive data in unapproved database responses, or help automate database monitoring and alerting policies. Due to the high potential speeds and transaction volumes in real-time database monitoring, these policies are usually limited to rules/patterns/categories. Vulnerability assessment policies may include more options because the performance demands are different.
- **Vulnerability Assessment:** Some vulnerability assessment tools can scan for basic DLP policy violations if they include the ability to passively monitor network traffic or scan storage.
- **Document Management Systems:** This is a common integration point for DLP solutions, but we don't see DLP included as a DMS feature.
- **Content Classification, Forensics, and Electronic Discovery:** These tools aren't dedicated to DLP, but we sometimes see them positioned as offering DLP features. They do offer content analysis, but usually not advanced techniques like partial document matching or database fingerprinting/matching.

Other Features and Integrations

The lists above include most of the DLP Light, feature, and integration options we have seen; but a few categories don't fit quite as neatly into our network/endpoint/storage divisions:

- **SIEM and Log Management:** All major SIEM tools can accept alerts from DLP solutions and correlate them with other collected activity. Some SIEM tools also offer DLP features, depending on what kinds of activity they can collect to perform content analysis on. Log management tools tend to be more passive, but increasingly include some similar basic DLP-like features when analyzing data. Most DLP users tend to stick with their DLP solutions for incident workflow, but we know of cases where alerts are sent to the SIEM for correlation or incident response, as well as others where the organization prefers to manage all security incidents in the SIEM.
- **Email Encryption:** Automatic encryption of email based on content was one of the very first third party integrations to appear on the market, and a variety of options are available. This is most frequently seen in financial and healthcare organizations (including insurance) with strict customer communication security requirements.

DLP Software as a Service (SaaS)

As described above, using a managed service for DLP is increasingly common given the staffing and technology requirements of an on-premise solution. There are a number of deployment models, including co-sourcing where you buy the DLP solution and have a service provider (sometimes the vendor) manage it for you. Another common model is to buy DLP as a service, where you pay a monthly fee for DLP capabilities.

Then the question becomes whether to run the console on-site or within your service provider's cloud. We don't subscribe to any religion about that decision, but make sure to fully understand your provider's multi-tenant architecture and how they isolate your data before committing to a cloud service.

Current DLP SaaS offerings fall into the following categories:

- **DLP for Email:** Many organizations are opting for SaaS-based email security rather than installing internal gateways. This is an effective and straightforward integration point for monitoring outbound email. Most services don't yet include full DLP analysis, but many major email security service providers have also acquired DLP solutions (sometimes before buying their email SaaS provider), so as you would expect extensive integration is available. This integration should enable policies and violations to synchronize from the cloud to your local management server.
- **DLP for Web Filtering:** As with email, we see organizations adopting cloud-based web content filtering to block web-based attacks before they hit the local network, and to better support remote users and locations. All content is already being scanned, so this is a nice fit for DLP SaaS. With similar acquisition to email services, we also hope to see integrated policy management and workflow for organizations obtaining their DLP web filtering from the same SaaS provider which supplies their on-premise DLP solution.
- **Full DLP:** Given the advance of cloud technologies, endpoint agents (and on-premise network devices) can now communicate to a central service in the cloud. Initially these services targeted smaller to mid-size organizations which didn't want the overhead of a full DLP solution and didn't have such deep requirements, but we increasingly see large implementations deployed as fully managed services running in a cloud environment.

Before jumping in with a SaaS provider, keep in mind that they won't be merely assessing and stopping external threats, but scanning for extremely sensitive content and policy violations. So we reiterate the importance of understanding a service provider's security posture, their multi-tenant architecture, and how they isolate your data from other customers.

Central Administration, Policy Management, and Workflow

DLP solutions use a central management server for administering enforcement and detection points, creating and administering policies, incident workflow, and reporting. These features are frequently the most important in the selection process, because processing alerts is how you benefit from a DLP system. There are many differences between products on the market, so rather than try to cover every possible feature we will focus on the baseline of most important functions.

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. The user interface must account for this mix of technical and non-technical staff, and 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 an 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 may be available only to authorized users or groups, which are typically kept in enterprise directories. The dashboard should focus on the information valuable to *this* user, not 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 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 will detail recommended functionality later in this paper.
- **Single Incident Display:** When a handler digs into an incident the display should cleanly and concisely summarize the reason for the alert, the user involved, criticality, severity (criticality is based on which policy is violated; severity on how much data is involved), related incidents, and all other information needed to make an informed incident disposition decision.

- **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. Policy creation and management is important enough that we'll cover it in detail later.

A DLP interface should be clean and easy to navigate. That may sound obvious but we're all far too familiar with poorly designed security tools which rely on the technical skills of the administrator to be usable. DLP is used outside Security — possibly even outside IT — so the user interface needs to work for a wide range of users with a variety of skill levels.

Management Functions

Hierarchical Management

DLP policies and enforcement often need to be tailored to the requirements 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 the 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 just 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 automatically to avoid accidentally tying a policy violation to an innocent user. For example a DLP product could associate a user to an IP address until the address is reassigned to another user. One company almost fired an employee because a contractor (not in Active Directory) was the next person to use that IP and committed a policy violation. Not knowing about the contractor, their 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 organizational structure.

Role-Based Administration

The system should 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. Your Active Directory might not reflect the user categories needed for monitoring and enforcement, so your DLP system should provide flexible support for monitoring and enforcement based on DLP specific groups and roles.

Policy Creation and Management

Policy creation and management is a critical function at the heart of DLP — and 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 nearly always requires technical skill.

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 which handlers and administrators can access the policy and violations. Not all policies are created equal, so each should also be assigned a sensitivity, with severity thresholds based on volume of violations. Each policy should be usable as a template for new policies, and the bundled policies associated with a given category should also be editable and available as templates for custom policies. Policy wizards are also useful, both for non-technical users and for experts quickly creating policies — such as a one-off to protect a single document.

Most users prefer 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.

Almost every policy needs some level of tuning, and good tools enable 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. It's also very helpful for a tool to test draft polices against recorded traffic.

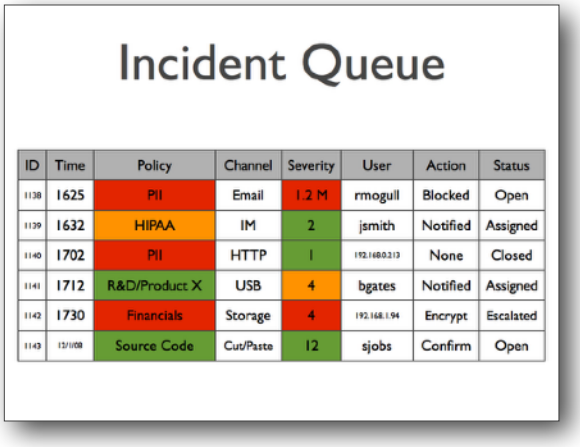
Policies include extremely sensitive information so they should be hashed, encrypted, or otherwise protected within the system. Some business units may have extremely sensitive policies which need to be protected against system administrators without explicit permission to see them. All policy violation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

Incident Workflow and Case Management

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

The first stop is the incident handling queue, a summary of all incidents either assigned to that handler, or unassigned but within the handler's enforcement domain. Incident status should be clearly indicated with color-coded sensitivity (based on the policy violated) and severity (based on volume of 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 — making it usable 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 shouldn't show up as 10 different incidents unless each attachment violates a different policy.



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 of other violations on 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 incidents and any supporting documentation, including time-stamps and data hashes. 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:** Of both the full system and the system configuration only, for migrations.
- **Import/Export:** for policies and violations. There should be some 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 prefer to completely manage the tool internally without requiring the administrator to learn or manage the platform. As much as possible, look for a DLP tools that lets you manage everything through this included interface.

Reporting varies widely across solutions; some use internal reporting interfaces while others rely on third-party tools. All tools ship with default reports, but if you need to create your own reports confirm your product can do what you need. Look for a mix of technical and non-technical reports, and if compliance is an issue consider tools that bundle compliance reports, which are ubiquitous among dedicated DLP solutions.

When you use storage or endpoint features you'll need a management interface that allows you to manage policies for servers, storage, and endpoints. The tool should support device grouping, performance and bandwidth management, rolling signature updates, and other features needed to manage large numbers of devices.

Beyond these basic features products differentiate themselves with other advances to 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.

DLP in the Cloud

It turns out a lot of organizations are using this cloud thing now, so they inevitably have questions about whether and how existing controls (including DLP) fit into the new world. As with most things related to cloud, there are significant differences in handling data leakage. But let's not put the cart before the horse. First we need to define what we mean by 'cloud,' with applicable use cases for DLP.

We could bust out the Cloud Security Alliance guidance and hit you over the head with a bunch of cloud definitions. But for our purposes it's sufficient to say that in terms of data access you are most likely dealing with:

- **SaaS:** Software as a Service is the new back office. That means whether you know about it or not, you have critical data in a SaaS environment, and it must be protected.
- **Cloud File Storage:** These services enable you to extend a device's file system to the cloud, replicating and syncing between devices and facilitating data sharing. Yes, these services are a specific subtype of SaaS (and PaaS, Platform as a Service), but the amount of critical data they hold, along with how differently they work than typical SaaS applications, demands that we treat them differently.
- **IaaS:** Infrastructure as a Service (IaaS) is the new data center. That means many of your critical applications (and data) will be moving to a cloud service provider — most likely Amazon Web Services, Microsoft Azure, or Google Cloud Platform. And inspection of data traversing a cloud-based application is, well… different, which that means protecting that data is also… different.

DLP is predicated on scanning data at rest and inspecting and enforcing policies on data in motion, which is a poor fit for IaaS. We don't really have endpoints suitable for DLP agent installation. Data is in either structured datastores such as databases or unstructured stores such as filesystem. Data protection for structured datastores defaults to application-centric methods, will unstructured cloud file systems are really just cloud file storage (which we will address later). So inserting DLP agents into an application stack isn't the most efficient or effective way to protect an application.

Compounding the problem, traditional network DLP don't fit IaaS well either. You have limited visibility into the cloud network; to inspect traffic, you would need to route it through an inspection point, which is likely to be expensive and/or lose key cloud advantages — particularly elasticity and anywhere access. Further, cloud network traffic is encrypted more often, so even with access to full traffic, inspection at scale presents serious implementation challenges.

So we will focus our cloud DLP discussion on SaaS and cloud file storage.

Cloud Versus Traditional Data Protection

The cloud is clearly different, but what exactly does that *mean*? If we boil it down to fundamentals, we still need to perform the same underlying functions — whether the data resides in a 20-year-old mainframe or the ether of a multi-cloud SaaS environment. To protect data we need to know where it is (discover), understand how it's being used (monitor), and then enforce policies to govern what is allowed and by whom — along with any additional necessary security controls (protect).

When looking at cloud DLP many users equate protection with encryption, but it's much more complicated than that, especially for SaaS. Managing keys across cloud and on-premise environments is significantly more complicated than before — you need to rely more on your provider, and architect data protection and encryption directly into your cloud technology stack.

Thinking about discovery, you remember the olden days — back as far as 7 years ago — when your critical data was either in your data centers or on devices you controlled? To be fair, even then it wasn't easy to find all your critical data, but at least you knew where to look. You could search all your file servers and databases for critical data, profile and/or fingerprint it, and then look for it across your devices and your network's egress points.

But as critical data started moving to SaaS applications and cloud file storage (sometimes embedded within SaaS apps), controlling data loss became more challenging, because it might leak without traversing a monitored egress point. We saw Cloud Access Security Brokers (CASB) emerge to figure out which cloud services were in use, so you could understand (kind of) where your critical data might be. At least you had a place to start looking.

Enforcement of data usage policies is also different in the cloud — you don't completely control SaaS apps, nor do you have an inspection/enforcement point on the network where you can look for sensitive data and prevent it from leaving. We keep hearing about lack of visibility in the cloud, and this is another case where it breaks classical security.

So what's the answer? It's in 3 letters you should be familiar with: A. P. I.

API Are Your Friends

Fortunately many SaaS apps and cloud file storage services provide APIs which allow you to interact with their environments, providing visibility and some degree of enforcement for your data protection policies. Many DLP offerings have integrated with leading SaaS and cloud file storage vendors to offer the ability to:

1. Know when files are uploaded to the cloud, and analyze them.
2. Know who is doing what with the files.
3. Encrypt or otherwise protect files.

With this access you don't need to see the data pass by, so long as the API reliably tells you new data has moved into the environment, with sufficient flexibility to monitor and manage it. The key to DLP in the cloud is integration with APIs for the services you use.

But what happens when you don't (or can't) get adequate integration with cloud environments via their API? You need to see the data somehow, and that's where a Cloud Access Security Broker comes into play.

Coexistence with CASB

CASB offers many functions, including visibility into cloud service usage within your environment. A CASB can also inspect traffic to cloud services by running it through a proxy. Of course this adds some overhead by routing traffic through the proxy, but the impact is highly dependent on the latency and response time requirements of the application and network architecture. Many CASB tools can also connect to cloud providers directly via their API to evaluate activity without a proxy. This requires the cloud provider to offer an adequate API, which is why proxy mode is often needed.

Because CASBs inspect traffic, vendors claim to provide DLP-like functions for traffic they see heading for cloud environments. Of course DLP on your CASB cannot provide visibility or enforcement for on-premise data. So deciding whether to look at the DLP Light capabilities on the CASB depends on whether you want or need consistent policy across both on-premise and cloud environments, or separate solutions for content monitoring are sufficient.

There is no right or wrong answer — your choice depends heavily on whether the policies you implement on internal networks map well enough to data moving to SaaS and cloud file storage.

Workflow Consistency

Once an alert triggers, where the data resides doesn't impact the processes internal folks use to verify the potential leak and assess the damage. So any workflow you have in place to handle data leakage should work wherever the data resides. Of course the tools for these processes differ, and your access to potentially compromised systems is radically different. For SaaS you simply have no access as a rule. Either way once you have a verified leak it is time for your incident response process.

Preventing data leaks in SaaS and cloud file storage can be very challenging. That said, as with most cloudy things, the place to start is by reassessing your processes and technologies to see what needs to change to be ready for the cloud.

But we know there will be more cloud use tomorrow than today, so the sooner you get your arms around protecting your content — regardless of where it resides — the better for your organization.

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 to use it, and the business processes around creating policies and managing incidents.

Define the Selection Team

Identify business units which 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 responsibility for enforcing controls on data. Content owners include business units which hold and use data. Content protectors tend to include departments such as Human Resources, IT Security, Legal, Compliance, and Risk Management. Once you identify the major stakeholders you'll want to bring them together for the next few steps.

Business Unit	Representative
IT Security	
CIO/IT Operations	
Legal	
Human Resources	
Risk Management	
Compliance	
Networking	
Email	
Storage	
Workstation/Endpoint	
Business Unit/Content Owner	
Business Unit/Content Owner	
Business Unit/Content Owner	

This list covers a superset of the people who tend to be involved with selection. Depending on the size of your organization you might need more or less, and in most cases the primary selection will be managed by 2-3 IT and IT security staff, but we suggest you include this larger group in the initial requirements generation process. The members of this team will also help obtain sample data for content analysis testing, and provide feedback on user interfaces and workflow if they will eventually be product users.

Stack Rank Your Data Protection Priorities and Define Data Types

The first step is to list out which major categories of data/content/information you want to protect. While it's important to be sufficiently specific for planning purposes, it's okay to stay fairly high-level. Definitions such as "PCI data," "engineering plans," and "customer lists" are good. Overly general categories like "corporate sensitive data" and "classified material" are too generic — they cannot be mapped to specific data types. This list must be prioritized — one good way to rank is to pull the business unit representatives together and force them to sort and agree to priorities, rather than having someone who isn't directly responsible (such as IT or security) determine rankings.

For each category of content listed in the first step, define the data category so you can map it to your content analysis requirements:

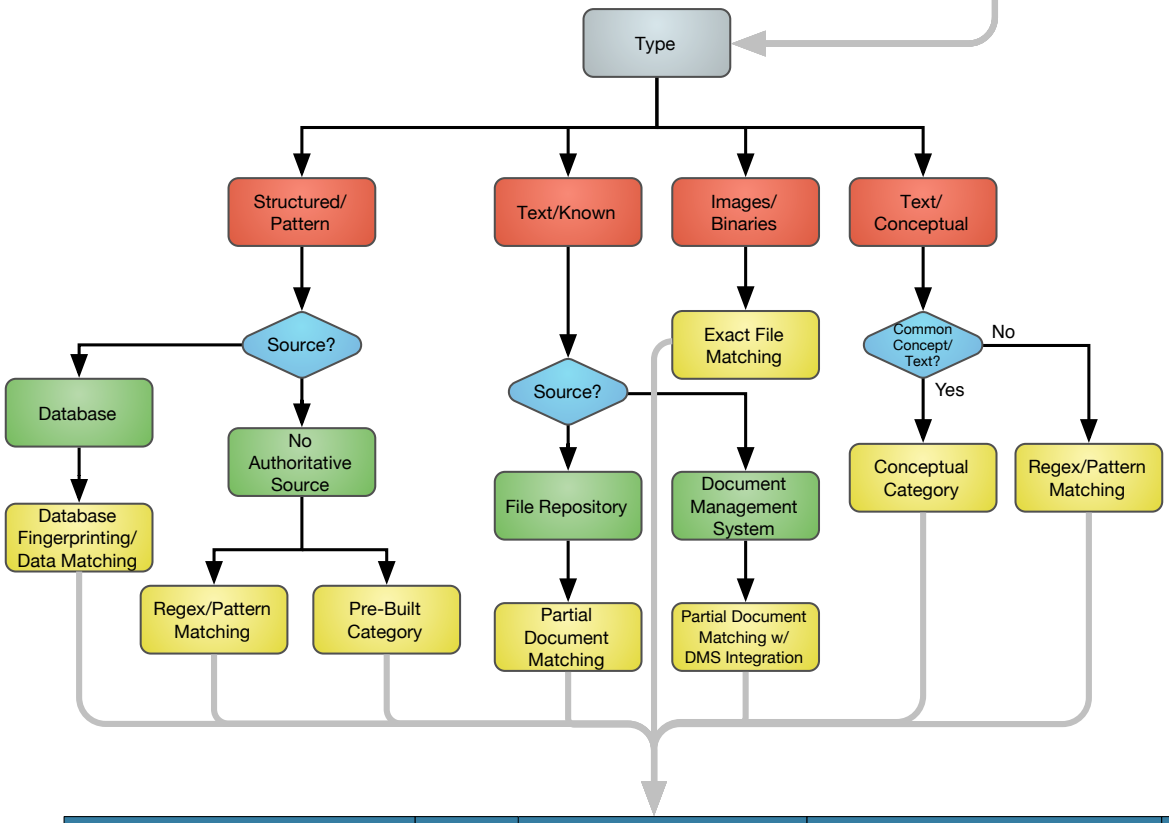
- Structured or patterned data includes credit card numbers, Social Security Numbers, and account numbers — it matches a defined pattern we can test for.
- Known text is unstructured content, typically found in documents, where we know the source and want to protect that specific information. Examples include engineering plans, source code, corporate financials, and customer lists.
- Images and binaries are non-text files such as music, video, photos, and compiled application code.
- Conceptual text is information that doesn't come from an authoritative source like a document repository but may contain keywords, phrases, or language patterns. This is pretty broad but examples include insider trading, job seeking, and sexual harassment.

Rank	Data/Content	Type
1		
2		
3		
4		
5		

Match Data Types to Required Content Analysis Techniques

Using the flowchart below you can evaluate required content analysis techniques based on data types and other environmental factors such as the existence of authoritative sources. This chart doesn't account for every possibility but is a good starting point and should define the high-level requirements for a majority of situations.

Rank	Data/Content	Type
1		
2		
3		
4		
5		



Content Analysis Techinque	Required	Additional Requirements	Requirements/Notes	Phase
Regex/Pattern Matching				
Pre-Built Categories		Categories? (*e.g.*, PCI/HIPAA)		
Database Fingerprinting/Data Matching		DB platform support		
		DB connection type support (ODBC vs. file extract)		
Partial Document Matching		Unusual document types?		
Partial Document Matching (DMS)		Document Management System platform support		
Conceptual Categories		Concept (*e.g.*, insider trading, job seeking)		
Exact File Matching				

Determine Additional Requirements

Some content analysis techniques include additional requirements, such as support for specific database platforms and document management systems. If you are considering database fingerprinting, determine whether you can work against live data in a production system or will rely on data extracts — periodic database dumps reduce performance overhead on the production system.

Define Rollout Phases

We haven't yet defined formal project phases but you should have an early idea whether each data protection requirement is immediate or something you can roll out later in the project. One reason is that many DLP projects are initiated to address some sort of breach or compliance deficiency relating to a single data type. This could lead to selecting a product based only on that requirement, which might mean problematic limitations down the road as you expand to protect other kinds of content.

Determine Monitoring/Alerting Requirements

Start by figuring out where you want to monitor your information: which network channels, storage platforms, and endpoint functions. High-level options include:

- Network
  - Email
  - Webmail
  - HTTP/FTP
  - HTTPS
  - Messaging/Collaboration
  - Generic TCP/IP
- Storage
  - File Shares
  - Document Management Systems
  - Databases
- Endpoint
  - Local Storage
  - Portable Storage
  - Network Communications
  - Copy & Paste
  - Print/Fax
  - Screenshots
  - Application Control

You may have additional requirements but these are the most common ones we encounter.

Determine Enforcement Requirements

As discussed already, most DLP tools offer various enforcement actions, which vary by channel and platform. The most basic enforcement option is 'Block': the activity is stopped when a policy violation is detected. An email might be filtered, a file not transferred to a USB drive, or an HTTP URL inaccessible. But most products also include other options, such as:

- **Encrypt:** Encrypt the file or email before allowing it to be sent or stored.
- **Quarantine:** Move the email or file into a quarantine queue for approval.
- **Shadow:** Allow a file to be moved to USB storage, but send a protected copy to the DLP server for later analysis.
- **Justify:** Warn the user that this action may violate policy, and require them to enter a business justification to store with the incident alert on the DLP server.
- **Change Rights:** Add or modify the file's Digital Rights Management.
- **Change Permissions:** Modify file permissions.

Location/Channel	Alert	Enforce
Network		
Email		
HTTP/FTP		
HTTPS		
IM/Messaging		
Generic TCP/IP		
Webmail		
Other		
Other		
Storage		
File Shares		
Document Management Systems		
Databases		
Other		
Endpoint		
Local Storage		
Portable Storage		
Network Communications		
Copy & Paste		
Print/Fax		
Screenshots		
Application Control		
Other		

Enforcement Actions	
Encrypt	
Block	
Quarantine	
Justify	

Enforcement Actions	
Encrypt	
Quarantine	
Change Rights	
Change Permissions	

Enforcement Actions	
Encrypt	
Block	
Shadow	

Map Content Analysis Techniques to Monitoring/Protection Requirements

As mentioned, DLP products vary in which policies they can enforce on which locations, channels, and platforms. Most often we see limitations on the types or size of policies that can be enforced on endpoints, including changes as the endpoint moves off and back onto the corporate network, because some polices and actions require communication with a DLP server.

For the final step in this part of the process, list your content analysis requirements for each monitoring/protection requirement you just defined. These directly translate to the RFP requirements at the core of most DLP projects: what you want to protect, where you need to protect it, and how.

Location/Channel	Content Analysis Techniques

Determine Infrastructure Integration Requirements

To work properly all DLP tools need some degree of integration with existing infrastructure. The most common integration points are:

- **Directory servers** to determine users and build user, role, and business unit policies. At minimum you need to know who to investigate when you receive an alert.
- **DHCP servers** so you can correlate IP addresses to users. You don't need this if all you are looking at is email or endpoints, but for other network monitoring it is critical.
- **SMTP gateways** — this can be as simple as adding your DLP tool as another hop in the MTA chain, but might be more involved.
- **Perimeter router/firewalls** for passive network monitoring. You need someplace to position the DLP sensor — typically a SPAN or mirror port, as discussed earlier.
- Your **web gateway** will probably integrate with your DLP system if you want to filter web traffic with DLP policies. If you want to monitor SSL traffic (you do!), you'll need to integrate with something capable of serving as a reverse proxy (Man in the Middle).
- **Storage platforms** to install an agent to scan your storage repositories directly, rather than relying on remote network/file share scanning.
- **Endpoint platforms** must be compatible with your endpoint DLP agent. You may want to use an existing software distribution tool to deploy it.

For each of these list out not only the platform you intend to integrate with, but any additional requirements, such as bandwidth or other performance requirements.

Infrastructure Component	Platform/Requirement
Network	
Directory Servers	
DHCP Servers	
Perimeter Router/Firewalls	
SMTP Gateways	
Email Servers	
Email Encryption System	
Web Gateways	
SSL/TLS Reverse Proxy	
IM/Messaging Gateways	
Other	
Other	
Storage	
File Servers	
Document Management Systems	
SharePoint	
Database Management Systems	
Digital Rights Management	
Other	
Other	
Endpoint	
Operating Systems	
Software Distribution/Update Tool	
Email Client	
Device Control Tool	
Remote Access Client	
DRM Client	
Other	
Other	
General	
SIEM	
Workflow Management	
Other	
Other	

We don't want to make this overly complex — many DLP deployments only integrate with a few of these infrastructure components, or the functionality is included within the main DLP platform. Integration might be as simple as plugging a DLP server into a SPAN port, pointing it at your directory server, and adding it into

the email MTA chain. But when developing requirements it is better to over-plan than to miss a crucial piece that blocks expansion later.

Finally, if you plan to deploy any database or document based policies, fill out the storage section of this table. Even if you don't plan to scan your storage repositories, you'll be using them to build document matching and database fingerprinting policies.

Determine Management, Workflow, and Reporting Requirements

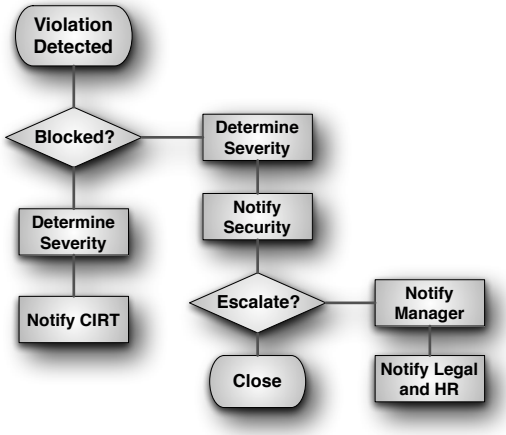
The last major set of requirements includes central management features, workflow, and reporting support. The following table includes common requirements, but is far from exhaustive:

Feature	Requirement
Management	
Consolidated UI for all DLP features	
Incident backup/restore	
Configuration backup/restore	
Policy backup/restore	
Hierarchical management	
Role-based management	
Per-role policy restrictions	
Policy creation wizards	
Regional policy support	
Endpoint agent management and performance tuning	
Storage repository management and performance tuning	
Business unit policy creation	
Automatic incident archiving	
Other	
Other	
Workflow	
Unified incident queue (network, endpoint, storage)	
Role-based incident handling (with regional and business unit support)	
Incident display restrictions (based on policy and sensitivity)	
Incident escalation and transfer	
Incident investigation (other similar violations or incidents by user)	
Internal case management	

Feature	Requirement
Per-handler incident queue display customization	
Email notification	
Non-technical UI support	
Organization dashboard	
Web-based interface	
Other	
Other	
Reporting	
Pre-defined compliance reports	
Additional pre-defined reports	
Internal reporting support	
Third-party reporting support	
Business/executive reports	
Scheduled email report generation	
Other	
Other	
Other	

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 policies? When a policy is violated what is the workflow to remediate it? When is HR notified? Legal? Who handles day-to-day policy violations? Is that a technical security role or non-technical, such as a compliance officer? The answers to these of questions will guide you toward different solutions for your workflow needs.



By the end of this phase you should have defined key stakeholders; convened a selection team; prioritized the data to protect; determined where you want to protect it; listed specific protection, integration, and management requirements; 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 requirements identified earlier are translated into specific technical features, and any additional requirements are considered. This is the time to come up with detailed criteria for directory integration, gateway integration, data storage, hierarchical deployment, endpoint integration, and so on that weren't already specified. 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 will develop a formal RFI (Request For Information) to release to vendors, and a rough RFP (Request For Proposals) you'll clean up and formally issue in the evaluation phase.

Evaluate Products

As with any product, it can be difficult to cut through the marketing to figure out whether a product really meets your needs. The following steps should minimize your risk and help you feel confident in your decision:

1. **Issue the RFI:** Larger organizations should issue an RFI though 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 DLP vendors which seem appropriate for your organization.
2. **Perform a paper evaluation:** Before bringing anyone in match any materials from vendors or other sources against your RFI and draft RFP. Your goal is to build a short list of 3 products which match your needs. 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 (if you are big enough), with a suite of basic rules. You'll want to 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 to any on-site demonstrations.
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 hard requirements such as directory integration. Then bring in any remaining products for in-house testing. To properly test products, place them on your network in passive monitoring mode and load up some sample rulesets representing the kinds of rules you want in production. This lets you compare products side by side, running equivalent rules, on the same traffic. If testing endpoint or storage support use a consistent set of endpoints or storage repositories. You'll also want to test any other features 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 your last chance to find problems during selection and before you pay. 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 its limits.
- Email and/or web proxy integration.
- Incident workflow: Review the working interface with 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 with your network and tune it. Do you need to filter traffic to reduce the amount of data going into the DLP appliance? Do you need to specify port and protocol combinations?
- Enforcement actions.

Summary

Procuring a Data Loss Prevention solution that works for your organization can be a long and arduous task, but understanding the capabilities of DLP and a structured selection process can help you choose an appropriate tool for your requirements.

After working with many organizations evaluating DLP we have a few conclusions. 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*.

Many of you probably hear that DLP is too complex to deploy, or generates too many false positives. This isn't the feedback we received from most of the people who actually deployed or managed DLP, but by the same token you are more likely to find Sasquatch than a security tool that's as easy to use and effective as a vendor's sales presentations. When performing *post-mortems* on struggling or failed DLP deployments, the typical culprits are a mixture of poorly structured implementations (including turning on a string of checklist policies and being overwhelmed with incidents) and improper expectations (such as being overwhelmed by high false positives from a DLP Light feature using a very crude regular expression rule). And there are always organizations which have bought and deployed DLP without bothering to figure out what content to protect, or who would be responsible for handling incidents. All you can do is shake your head.

The key to a successful DLP deployment is knowing your needs, understanding the capabilities of your tool, and properly setting expectations. Know what you want to protect, how you want to protect it, and where you need to integrate with existing infrastructure before you let a vendor in the door.

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 the wrong people are seeing policy violations, or that your new purchase isn't capable of protecting the sensitive data of a business unit not included in your selection process.

DLP products provide very high value for organizations which plan properly and understand how to take full advantage of them. Focus on the features 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.

About the Analyst

Rich Mogull, Analyst and CEO

Rich has twenty years of experience in information security, physical security, and risk management. He specializes in data security, application security, emerging security technologies, and security management. Prior to founding Securosis, Rich was a Research Vice President at Gartner on the security team where he also served as research co-chair for the Gartner Security Summit. Prior to his seven years at Gartner, Rich worked as an independent consultant, web application developer, software development manager at the University of Colorado, and systems and network administrator. Rich is the Security Editor of TidBITS, a monthly columnist for Dark Reading, and a frequent contributor to publications ranging from Information Security Magazine to Macworld. He is a frequent industry speaker at events including the RSA Security Conference and DefCon, and has spoken on every continent except Antarctica (where he's happy to speak for free — assuming travel is covered).

Mike Rothman, Analyst/President

Mike's bold perspectives and irreverent style are invaluable as companies determine effective strategies to grapple with the dynamic security threatscape. Mike specializes in the sexy aspects of security — such as protecting networks and endpoints, security management, and compliance. Mike is one of the most sought-after speakers and commentators in the security business, and brings a deep background in information security. After 20 years in and around security, he's one of the guys who "knows where the bodies are buried" in the space.

Starting his career as a programmer and networking consultant, Mike joined META Group in 1993 and spearheaded META's initial foray into information security research. Mike left META in 1998 to found SHYM Technology, a pioneer in the PKI software market, and then held executive roles at CipherTrust and TruSecure. After getting fed up with vendor life, Mike started Security Incite in 2006 to provide a voice of reason in an over-hyped yet underwhelming security industry. After taking a short detour as Senior VP, Strategy at eIQnetworks to chase shiny objects in security and compliance management, Mike joined Securosis with a rejuvenated cynicism about the state of security and what it takes to survive as a security professional.

Mike published The Pragmatic CSO <http://www.pragmaticcso.com> in 2007 to introduce technically oriented security professionals to the nuances of what is required to be a senior security professional. He also possesses a very expensive engineering degree in Operations Research and Industrial Engineering from Cornell University. His folks are overjoyed that he uses literally zero percent of his education on a daily basis. He can be reached at mrothman (at) securosis (dot) com.

About Securosis

Securosis, LLC is an independent research and analysis firm dedicated to thought leadership, objectivity, and transparency. Our analysts have all held executive level positions and are dedicated to providing high-value, pragmatic advisory services. Our services include:

- **Primary research publishing**: We publish the vast majority of our research for free through our blog, and package the research as papers that can be licensed for distribution on an annual basis. All published materials and presentations meet our strict objectivity requirements, and follow our Totally Transparent Research policy.
- **Cloud Security Project Accelerators**: Securosis Project Accelerators (SPA) are packaged consulting offerings to bring our applied research and battle-tested field experiences to your cloud deployments. These in-depth programs combine assessment, tailored workshops, and ongoing support to ensure you can secure your cloud projects better and faster. They are designed to cut months or years off your projects while integrating leading-edge cloud security practices into your existing operations.
- **Cloud Security Training:** We are the team that built the Cloud Security Alliance CCSK training class and our own Advanced Cloud Security and Applied SecDevOps program. Attend one of our public classes or bring us in for a private, customized experience.
- **Advisory services for vendors**: We offer a number of advisory services to help our vendor clients bring the right product/service to market in the right way to hit on critical market requirements. Securosis is known for telling our clients what they NEED to hear, not what they want to hear. Clients typically start with a strategy day engagement, and then can engage with us on a retainer basis for ongoing support. Services available as part of our advisory services include market and product analysis and strategy, technology roadmap guidance, competitive strategies, etc. Though keep in mind, we maintain our strict objectivity and confidentiality requirements on all engagements.
- **Custom Research, Speaking and Advisory**: Need a custom research report on a new technology or security issue? A highly-rated speaker for an internal or public security event? An outside expert for a merger or acquisition due diligence? An expert to evaluate your security strategy, identify gaps, and build a roadmap forward? These defined projects bridge the gap when you need more than a strategy day but less than a long-term consulting engagement.

Our clients range from stealth startups to some of the best known technology vendors and end users. Clients include large financial institutions, institutional investors, mid-sized enterprises, and major security vendors. For more information about Securosis, visit our website: <http://securosis.com/>.