Securosis, L.L.C.

DLP Content Discovery: Best Practices for Stored Data Discovery and Protection

by Rich Mogull

This Report Sponsored by:

Author's Note

The content in this report was developed independently of any sponsors. It is based on material originally posted on the Securosis blog but has been enhanced and professionally edited.

This report is sponsored by Symantec Inc.

Special thanks to Chris Pepper for editing and content support.

Sponsored by Symantec

Vontu, now part of Symantec, is the leading provider of Data Loss Prevention solutions that combine endpoint and network-based technology to accurately detect and automatically protect confidential data wherever it is stored or used. By reducing the risk of data loss, Vontu solutions from Symantec help organizations ensure public confidence, demonstrate compliance and maintain competitive advantage. Vontu Data Loss Prevention customers include many of the world's largest and most data-driven enterprises and government agencies. Vontu products have received numerous awards, including IDG's InfoWorld 2007 Technology of the Year Award for "Best Data Leak Prevention," as well as SC Magazine's 2006 U.S. Excellence Award for "Best Enterprise Security Solution" and Global Award for "Best New Security Solution." For more information, please visit http://go.symantec.com/vontu.

Copyright

Table of Contents

Introduction

Intelligent Information Risk Reduction

The modern enterprise is a veritable ocean of unmanaged information. Despite many years, and many dollars, invested in gaining some control over the massive volumes of content generated by our employees, customers, and business operations, few organizations have a clear idea where their sensitive information is located, or how it's secured. While some content happens to reside in protected servers, document management systems, and databases, far more is scattered across email files, endpoints, and completely unknown file shares. Enterprises are at risk of compliance violations, insider abuse, and external attacks — not because they can't protect their sensitive data in known locations, but because they have very little understanding of where the content is really located, how it's protected, and how it's being used.

One of the most promising techniques to help reduce this risk is labelled Data Loss Prevention (DLP). While most people think of network monitors when they hear "DLP", the truth is that DLP tools are often as valuable when used to protect data at rest, rather than only data in motion.

Let's review our definition of DLP:

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

Content Discovery, the ability to scan and monitor data at rest, is one of the most important features of any DLP solution — one with significant potential to reduce enterprise risk. While network DLP manages how users communicate sensitive information, content discovery documents where sensitive information is stored within the enterprise, and often how it's used. Content discovery is likely more effective at reducing enterprise risks than network monitoring, alone and is one reason organizations should consider full DLP suites over single-channel products.

Why?

Consider the potential of knowing nearly every location where you store sensitive information, based on deep content analysis, and also who has access to that data. Of being able to continuously monitor your environment and receive notification when sensitive content is moved to unapproved locations, or even if its access rights are changed. Of, in some cases, being able to proactively protect the content by quarantining, encrypting, or moving it when policy violations occur.

Content discovery, by providing deep insight into the storage and use of your sensitive information, is a powerful risk reduction tool — one that often also reduces audit costs.

DLP Content Discovery in Action

Before we jump into a technology description, let's highlight a few simple use cases to demonstrate this risk reduction:

- Company A creates a policy to scan their storage systems for unencrypted credit card numbers. They provide this report to their PCI auditor to reduce audit costs and prove they are not storing cardholder information in violation of policy.
- Company B is developing a new product. They create a policy to generate an alert if engineering plans appear anywhere except on protected servers.
- Company C, a software development company, uses their discovery tool to ensure that source code only resides in their source code management repository. They scan developer systems to prevent source code from being stored outside the approved development environment.
- Company D, an insurance company, scans employee laptops to ensure employees don't store medical records to work on at home, and only access them through the company's secure web portal.

In each case we're not talking about preventing a malicious attack, although we are making it a bit harder for an attacker to find anything of value; we're instead focused on reducing risk by reducing exposure and gaining information on the use of content. Sometimes it's for compliance, sometimes it's to protect corporate intellectual property, and other times it's simply to monitor internal compliance with corporate policies.

According to an informal survey of major DLP vendors, content discovery is a part of 50-60% of initial DLP purchases, with deployment within the first 12 months of implementation.

As with most of our security tools, content discovery isn't perfect. Monitoring isn't always in real time, and it's possible to miss some storage locations, but even without perfection we can materially reduce enterprise risks.

Technology Overview

Broad Scanning with Deep Content Analysis

As a DLP tool, content discovery's key distinguishing feature is deep content analysis based on central policies. This contrasts with non-DLP data-at-rest discovery tools, such as content classification and e-discovery. While covering all content analysis techniques is beyond the scope of this report, some of the more common ones include partial document matching, database fingerprinting (or exact data matching), rules-based, conceptual, statistical, predefined categories (like PCI compliance), and combinations of the above. They offer far deeper analysis than just simple keyword and regular expression matching. Ideally, DLP content discovery should also offer preventative controls, not just policy alerts after violations occur. How does all this work?

Architecture

At the heart is the central policy server; the same system that manages the rest of your DLP deployment. The three key features of the central management server are policy creation, deployment management, and incident handling (workflow). Large deployments may have multiple central servers, but they all interconnect hierarchically.

Data at rest is analyzed using one of four techniques, each embodied in its own scanning component:

- *Remote scanning*: Either the central policy server or a dedicated scanning server to access storage repositories via network shares or other administrative access. Files are scanned for content violations. Connections are often made using administrative credentials, and any content transfered 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 the storage to increase speed and limit network impact. It supports scanning nearly any storage repository, but even with optimization remote scanning performance is limited by the network.
- *Server agent*: A lightweight agent is installed on the server and scans content locally. Agents can be tuned to limit performance impact, and results are sent securely to the central management server. While scanning performance is higher than remote scanning, this requires platform support and local installation.
- *Endpoint agent*: While you can scan endpoints (workstations) remotely using administrative file shares, that eats network bandwidth. DLP solutions increasingly include endpoint agents with local discovery capabilities. These agents normally provide additional DLP functions, such as USB monitoring and/or blocking.
- *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 possibly enforce actions within the system.

A good content discovery tool will understand file context, not just content. For example, some tools can analyze access controls on files and use the corporate directory to understand which users and groups have what access. Thus the

accounting department can access corporate financials, but any such files with inadequate access controls are identified for remediation. Engineering teams can see engineering plans, but access controls are automatically updated to prevent access by the accounting team if engineering content shows up in the wrong repository.

From an architectural perspective you'll want to look for solutions which support multiple options, with performance that meets your requirements.

Enforcement actions

Once the DLP discovery tool determines something is out of place, it can (depending on the tool) take enforcement actions ranging from alerts to active protection, or some combination. In cases where files are restricted, moved, or encrypted, an unprotected plain text file can be dropped into the same location to notify users who to contact with questions now that they can't access the file.

- *Alert*: An alert is recorded as a DLP incident. This is the base action, triggered no matter what else occurs.
- *Notify*: Email is sent to either the content owner (based on access controls and directory integration), the policy manager (based on the DLP policy), or pretty much anyone else (such as the content owner's manager).
- *Restrict access:* Access controls are modified to restrict access, *e.g.,* block anyone except a security administrator from accessing the file so it's protected until the violation can be reviewed.
- *Move/quarantine*: The file is moved to a secure repository.
- *Encrypt*: The file is encrypted. It could be protected with a generic corporate key, or something more specific such as a group, security, or administrative key.

Management

Ideally your content discovery capabilities should be managed through the same server as the rest of your DLP deployment. This helps maintain consistent policies, workflow, and incident handling. Here are a few discovery-specific capabilities to look for:

- *Policy creation*: Data at rest policies should be completely integrated with your other DLP policies. This enables you to define a type of content only once (*e.g.,* credit card numbers or engineering plans), and then apply appropriate alerting and protection rules at rest, in use, and in motion as aspects of a single policy. Policies should allow for fine-grained control based on user groups through directory integration.
- *Directory integration*: All DLP solutions can identify IP and email addresses, but for content discovery they also need to understand network users and groups, to tie into access controls.
- *Repository management*: This is the part of the system where you identify and group storage repositories such as servers, shares, and document management systems. For crawling, it's where you store access credentials, and for agents this includes agent management. Ideally you should be able to tag groups of repositories to make policy building easier (*e.g.,* "accounting" or "engineering"). This is the place to set scanning frequency, schedule, bandwidth/performance throttling, incremental vs. complete scans, and other basic functional preferences.

Workflow and Reporting

Workflow should be completely integrated into the base DLP incident handling queue. Content discovery related incidents should appear right alongside in-motion and in-use incidents, although you might assign a different incident handler for at-rest policies, depending on organizational needs. For example, you could assign a specific handler for all storage-related PCI violations, while keeping network violations in the general queue. If you encrypt, quarantine, or otherwise protect files, the DLP solutions must include management of those controls so you can release and restore as needed.

Reporting, on the other hand, should include content discovery specific reports, especially audit reports to facilitate compliance. While a report on all transmission of credit card numbers via email may not be the kind of thing you want to send an auditor, a report showing that you don't have any unprotected numbers in any known storage location would be more useful. Also look for the ability to generate reports for business unit managers, storage administrators, audit/legal/compliance, and other non-technical personnel. Because scans are run periodically, the solution should allow you to automatically schedule and distribute reports, rather than requiring them to be run manually each time.

Deployment Best Practices

Preparing for Deployment

Before installing a DLP tool and creating any policies, first focus on setting expectations, prioritizing, and defining your internal processes. The greatest barrier to successful deployment isn't any technology issue, but rather failure of the enterprise to understand what to protect, decide how to protect it, and recognize what's reasonable in a real-world environment.

Setting Expectations

The single most important factor for any successful DLP deployment — content discovery or otherwise — is properly setting expectations at the start of the project. DLP tools are powerful, but far from a magic bullet or black box that can make all data completely secure. When setting expectations you need to pull key stakeholders together in a single room and define what's achievable with your solution. All discussion at this point assumes you've already selected a tool. Some of these practices deliberately overlap steps from a typical selection process, since by this point you'll be able to refine those earlier decisions, based upon a much clearer understanding of the capabilities of your chosen tool.

In this phase, discuss and define the following:

- What kinds of content you can protect, based on the content analysis capabilities of your tool.
- Expected accuracy rates for those different kinds of content; for example, you'll have a much higher false positive rate with statistical or conceptual techniques than partial document or database matching.
- Protection options: Can you encrypt? Move files? Change access controls?
- Performance, based on scanning techniques.
- How much infrastructure you'd like to cover (which servers, endpoints, and other storage repositories).
- Scanning frequency (days? hours? near-continuous?).
- Reporting and workflow capabilities.

It is extremely important to start with a phased implementation. It's completely unrealistic to expect to monitor every nook and cranny of your infrastructure with your initial rollout. Nearly every organization finds they are more successful with a controlled, staged rollout that slowly expands breadth of coverage and types of content to protect.

Prioritization

If you haven't already prioritized your information during the selection process, you need to get all major stakeholders together (business units, legal, compliance, security, IT, HR, etc.) and agree upon which kinds of information are most important, and which to protect first. We recommend you first rank major information types (*e.g.,* customer PII, employee PII, engineering plans, corporate financials, etc.), then re-order them based on priorities for monitoring and protecting within your DLP content discovery tool.

In an ideal world your prioritization should directly match the order of protection, but while some data might be more important to the organization (engineering plans) other data may need to be protected first due to exposure or regulatory requirements (PII). You may also need to tweak the order based on the capabilities of your tool.

After your prioritize information types to protect, run through and determine approximate timelines for deploying content policies for each type. Be realistic, and understand that you'll need to both tune new policies and leave time for the organization to become comfortable with any required business changes.

We'll look further at how to roll out policies and what to expect in terms of deployment times later.

Defining Processes

DLP tools are, by their very nature, intrusive. Not in terms of breaking things, but rather in terms of the depth and breadth of what they find. Organizations are strongly advised to define their business processes for dealing with DLP policy creation and violations before turning on the tools. Here's a sample process for defining new policies:

1. Business unit requests policy from DLP team to protect specific content type.
2. DLP team meets with business unit to determine goals and protection requirements.
3. DLP team engages with legal/compliance to determine any legal or contractual requirements or limitations.
4. DLP team defines draft policy.
5. Draft policy tested in monitoring (alert only) mode without full workflow; then tuned to acceptable accuracy.
6. DLP team defines workflow for new policy.
7. DLP team reviews final policy and workflow with business unit to confirm needs have been met.
8. Appropriate business units notified of new policy and any required changes in business processes.
9. Policy deployed in production environment in monitoring mode, but with full workflow enabled.
10. Protection and workflow approved.
11. Protection and enforcement actions enabled.

And here's one for policy violations:

1. Violation detected; appears in incident handling queue.
2. Incident handler confirms incident and severity.
3. If action required, incident handler escalates and opens investigation.
4. Notify business unit contact assigned to triggered policy.
5. Evaluate incident.
6. Take protective actions.
7. If file moved/protected, notify user and deposit placeholder file with contact information.
8. Notify employee manager and HR, who may take corrective actions.
9. Perform required employee education.
10. Close incident.

These are, of course, just basic examples, but they should give you a good idea of where to start. You'll notice that these depend heavily on knowing who your users are, their job roles, and who they report to. Before deployment, it's important to evaluate your directory infrastructure to know if it accurately reflects your organizational structure. If not, consider updating your directory services or adjusting your processes to account for the manual identification of users in your incident workflow.

Deployment

By this point you should know what policies you'd like to deploy, what content to start protecting, how you'd like to grow that protection after initial deployment, and your workflow for policy violations.

Now we're ready to move beyond planning into deployment.

- *Integrate with your infrastructure*: DLP content discovery tools require either a local agent or server access to scan content in a repository. In this stage you define the initial repositories to scan and either install agents or load credentials into the DLP system. For endpoints, you could scan C$ or D$ (administrative access to all files) remotely, but a local agent is your best option for managed systems. If you haven't already, you also need to integrate with your enterprise directory servers so the DLP tool understands users, groups, and roles.
- *Build initial policies*: For your first deployment, you should start with a small subset of policies, or even a single policy, in alert or content classification mode (where the tool reports on sensitive data, but doesn't generate policy violation alerts).
- *Baseline, then expand deployment*: Deploy your initial policies on a limited number of storage repositories or endpoints. Once you have a good feel for the effectiveness of the policies, performance, and enterprise integration, you can expand into a wider deployment, covering more of the enterprise. After the first few rounds you'll have a good understanding of how quickly, and how widely, you can roll out new policies.
- *Tune policies*: Even stable policies may require tuning over time. In some cases it's to improve effectiveness, in others to reduce false positives, and in still other cases to adapt to evolving business needs. You'll want to initially tune policies during baselining, but continue to tune them as the deployment expands. Most DLP clients report that they don't spend much time tuning policies after baselining, but it's always a good idea to keep your policies current with enterprise needs.
- *Add enforcement and protection*: By this point you should understand the effectiveness of your policies, and have educated users where you found policy violations. You can now start switching to enforcement or protective actions, such as moving, encrypting, and changing access controls on files. Any time you make a file inaccessible, you should leave a plain-text contact note (or send the user an email) so they know why the file is missing and how to request an exception. If you're making a major change to established business process (*e.g.,* restricting access to a common content type to meet a new compliance need), consider scaling out enforcement options on a business unit by business unit basis. We do see some cases where users deploy enforcement actions earlier in the process; typically when cleaning up old data or with well understood policies on new areas of the infrastructure.

Deploying DLP content discovery isn't really very difficult; the most common mistake enterprises make is applying policies too quickly and too broadly.

It's also important to keep in mind that there are four general types of discovery deployments. With a monitoring and alerting deployment you roll out and generate alerts in the DLP system, which are then followed up on by incident handlers. These deployments are often for sensitive data types where you don't want immediate protection, but do want to kick off corrective action or user education. The second type of deployment is where you add content protection actions, such as encryption. It's typically for very sensitive data types, and as we've outlined above often follows an alerting-only deployment. In some cases content protection is used for content cleaning, and enforcement actions are deployed earlier. In a compliance deployment we scan for selective data related to regulatory compliance, such as credit card numbers — both to ensure sensitive data remains within appropriate containers, and also to generate compliance reports to show auditors that content is being handled appropriately, or that the organization knows where it’s located.

The last deployment model is completely different — content classification. In this case you scan with a very wide scope, often using general policies, to identify and classify systems based on the content they hold. Or in some cases, you might tag the content as part of a broader classification initiative. Content classification deployments aren't concerned with alerts or enforcement actions, but rather use the tools to classify systems and content.

Use Cases

We've finished our review of DLP content discovery best practices, as well as how to deploy and maintain a system. Now we'll focus on a couple use cases that illustrate how it all works together. These are synthetic case studies, based on interviews with real DLP customers.

DLP Content Discovery for Risk Reduction and to Support PCI Audit

RetailSportsCo is a mid-sized online and brick-and-mortar sporting goods retailer, with about 4,000 headquarters employees and another 2,000 retail employees across 50 locations. They classify as a Level 2 PCI merchant due to their credit card transaction volume and are currently PCI compliant; they struggled through the process and ended up having a series of compensating controls approved by their auditor, but only for their first year.

During the audit it was discovered that credit card information had proliferated throughout the organization. It was scattered through hundreds of files on dozens of servers; mostly Excel spreadsheets and Access databases used (and later ignored) by different business units. Since storage of unencrypted credit card numbers is prohibited by PCI, their auditor required them to remove or secure these files. Audit costs for the first year increased significantly due to the time spent by the auditor confirming that the information was destroyed or secured.

RetailSportsCo purchased a DLP solution and created a discovery policy to locate credit card information across all storage repositories and employee systems. The policy was initially deployed against the customer relations business unit servers, where over 75 files containing credit card numbers were discovered. After consultation with the manager of the department and employee notification, the tool was switched into enforcement mode and all these files were quarantined into an encrypted repository.

In phase 2 of the project, DLP endpoint agents were installed on the laptops of sales and customer relations employees (about 100 people). Users and managers were educated, and the tool discovered and removed approximately 150 additional files. Phase 3 added coverage of all known storage repositories at corporate headquarters. Phase 4 expanded scanning to storage at retail locations, over a period of 5 months. The final phase will add coverage of all employee systems in the first few months of the coming year, leveraging their workstation configuration management system for a scaled deployment.

Audit reports were generated to show exactly which systems were scanned, what was found, and how it was removed or protected. Their auditor accepted the report, which reduced audit time and costs materially (more than the total cost of the DLP solution). One goal of the project is to scan the entire enterprise at least once per quarter, with critical systems scanned on either a daily or weekly basis. RetailSportsCo has improved security and reduced risk by reducing the number of potential targets, and reduced compliance costs by being able to provide auditors with acceptable reports demonstrating compliance.

DLP Content Discovery to Reduce Competitive Risk (Industrial Espionage)

EngineeringCo is a large high-technology manufacturer of consumer goods with 51,000 employees. In the past they've suffered from industrial espionage, when the engineering plans for new and existing products were stolen. They also suffered a rash of unintentional exposures and product plans have been accidentally placed in public locations, including the corporate web site.

EngineeringCo acquired a DLP content discovery solution to reduce these risks and protect their intellectual property. Their initial goal was to reduce the risk of exposure of engineering and product plans. Unlike RetailSportsCo, they decided to start with endpoints, then move into scanning enterprise storage repositories. Since copies of all engineering and product plans reside in the enterprise content management system, they chose a DLP solution that could integrate with their system and continuously monitor selected locations and automatically build partial-document matching policies for all documents in the system. The policy was tested and refined to ignore common language in the files, such as corporate headers and footers, which initially caused every document using the corporate template to register in the DLP tool.

EngineeringCo started with a phased deployment to install their DLP endpoint discovery agent on all corporate systems. In phase 1, the tool was rolled out to 100 systems per week, starting with product development teams. The first phase was delayed slightly as they realized they needed to update the roles and users in their Active Directory servers to more accurately reflect their organizational structure and who was on what teams. The initial policy allowed those teams access to the sensitive information, but documented what was on their systems. Those reports were later mated to their encryption tool to ensure that no unencrypted laptops hold the sensitive data. Phase 2 expanded deployment to the broader enterprise, initially in alerting mode. After 90 days the product was switched into enforcement mode and all identified content outside the product development teams was quarantined with an alert sent to the user, who could request an exemption. Initial alert rates were high, but user education reduced levels to only a dozen or so "violations" a week by the end of the 90-day grace period.

In the coming year EngineeringCo plans to refine their policy to prevent product development employees from placing registered documents onto portable storage. The network component of their DLP tool already restricts emailing and other file transfers outside of the enterprise. They also plan on adding policies to protect employee healthcare information and customer account information.

Conclusion

These are, of course, fictional best practices examples, but they're drawn from discussions with dozens of DLP clients. The key takeaways are:

- Start small, with a few simple policies and a limited scanning footprint.
- Grow deployments as you reduce incidents and violations to keep your incident queue under control and educate employees.
- Start with monitoring and alerting and employee education, then move on to enforcement.
- This is risk reduction, not risk elimination. Use the tool to identify and reduce exposures but don't expect it to magically solve all your data security problems.
- When you add new policies, test first with a limited audience before rolling out to the full scope, even if you are already covering the entire enterprise with other policies.

DLP content discovery is a flexible tool offering a myriad of risk reduction and content classification benefits. From providing high level overview of the use of sensitive information in your organization, to reducing compliance costs, to proactively enforcing controls on data, it's a powerful technique for risk reduction.

About the Author

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

About Securosis

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

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

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

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