Likewise®

Whitepaper

Securing Unstructured Data

Protecting Sensitive Files by Uniting Identity, Security, and Storage

Table of Contents

By Steve Hoenisch,
Likewise Software

Executive Summary

Unstructured data is growing rapidly — and up to 40 percent of it contains sensitive information. Companies that identify and consolidate their sensitive unstructured data and then establish consistent policies to protect it can reduce operational costs and security risks.

At the same time, harnessing the big explosion in unstructured data can lead to new applications, new lines of business, and new ways to maximize revenue.

To lay the foundation for securing and harnessing sensitive unstructured data, several requirements must come together: cross-platform storage, consistent security policies and access control, object tracking, event monitoring, and secure auditing and reporting.

Consolidating unstructured data housed in silos into cross-platform, multiprotocol file servers or NAS systems sets the stage for a common security model. A single security system to manage identities and control access enforces consistent data-security policies, including those for file servers. Meantime, event monitoring and auditing that is close to the data tracks access patterns and logs changes to sensitive files. Secure reporting helps demonstrate regulatory compliance and show chain of custody over tracked folders and files.

By uniting storage, identity, and security, you can protect sensitive unstructured data while making it highly available to those who need it so they can become more efficient, innovate, and create value for competitive advantage.

Introduction

Unstructured data is rapidly piling up in file systems at an alarming rate. According to industry analysts, unstructured data is growing faster than all other types of data and will increase by as much as 800 percent during the next five years. Analysts are beginning to refer to it as the big-data explosion.

Unstructured data, however, is not innocuous. It often contains sensitive information, which if exposed could damage your reputation and lead to legal fees and fines. Respondents to an industry survey by the Aberdeen Group estimated that 40 percent of their sensitive data is in unstructured formats such as PDF and Microsoft Office files like Word and Excel.

“Secrets tend to be unstructured. Although there are exceptions, intellectual property that provides longer-term advantage isn’t neatly codified into rows and columns — they are more messily and abstractly described in Word documents, embedded in presentations, and enshrined in application-specific formats,” Forrester analyst Andrew Jaquith says in Selecting Data Protection Technologies.

Because security policies and rules governing access are often inadequate in the face of the massive growth of unstructured data, sensitive information stored on collaboration systems, file servers, and network attached storage (NAS) devices is in danger of compromise.

At the same time, protecting unstructured data must be balanced against making the data rapidly and easily available to foster efficient business operations, value creation, and competitive advantage. There is a tradeoff between protection on the one hand and enabling business operations and innovation on the other. To be successful, a balance must be struck between security and availability.

The need for highly scalable file storage is a ubiquitous necessity for organizations from government offices and university departments to small companies and large corporations — all of which are generating vast amounts of unstructured data at an ever-increasing rate. Since 40 percent of that unstructured data inevitably contains sensitive information, the explosion of unstructured data poses a significant security risk and a major compliance problem.

The complexity of heterogenous storage systems further complicates the issue. Within an organization, some departments inevitably use Windows computers and CIFS-based file servers while others use Unix or Linux systems and NFS-based storage. “Complexity, too much heterogeneity, and duplication of systems make IT too expensive,” Forrester Research analyst Galen Schreck says in Assessing Your IT Infrastructure Architecture.

NFS-based file servers for Unix users and CIFS-based file servers for Windows users, combined with the inability to interoperate between the two, turn cross-platform storage into a complicated network of mixed systems containing duplicated unstructured data — systems that are difficult to secure with a common directory service and consistent security policies.

Securing the unstructured data on those file systems is, for most businesses, an operational and legal imperative driven by the need to reduce costs, minimize security risks, and comply with regulations. “Management of business content is now a legal imperative in many industries. With the upsurge in modes of digital communication, most companies are not prepared. Only 20% report that they’re very confident that if challenged, they could demonstrate that their digital information is accurate, trustworthy, and accessible. In a world where mismanagement of information can lead to reputation risk, privacy violations, and regulatory issues, proper management of business content and communications is no longer a luxury,” Rob Koplowitz and Ted Schadler of Forrester Research write in Benchmarking Your Collaboration Strategy.

This white paper describes a number of problems that make it difficult to protect unstructured data and then highlights some technologies and policies that you should consider to help secure your organization’s mass of sensitive unstructured files while reducing operational costs.

Problems in Protecting Unstructured Data

Although the issues related to protecting unstructured data are multifaceted because of its ubiquitous nature, they typically evolve around the following fundamental problems:

Unstructured data is widely distributed and segmented by silos.

Because of distributed environments, disparate operating systems, and different protocols, unstructured data ends up all over the place — but frequently in file servers segmented by protocol. Unix data ends up in NFS-based storage devices, and Windows data ends up in CIFS-based storage devices. Data silos — stored data segmented by protocol — hinders cost-saving storage consolidation, a common security model, risk-averting access controls, and mandatory compliance.

Unstructured data is hard to locate and classify.

Distributed systems and data silos make unstructured data difficult not only to locate but also to classify. Locating and then classifying the data based on sensitivity, risk, compliance, or other categories is an important step toward being able to protect it.

Decentralized access control results in a security model that is not uniform and in security policies that are not consistent.

Many organizations continue to use multiple authentication and access control systems, such as NIS or LDAP for Unix machines and Active Directory for Windows computers. There is frequently one access control model in place for CIFS-based storage and another for NFS-based file servers.

Decentralized access control makes it difficult to implement uniform, consistent security policies that control access to unstructured data. Putting in place a uniform security model and using a common identity management system to enforce consistent data-security policies addresses not only who can access what data but also what actions may be taken on the data, such as modifying or deleting it.

Determining who can access what is difficult.

Unstructured data that is distributed across the enterprise, segmented into storage silos, and controlled by disparate access control systems compounds the problem of determining the users and groups who can access sensitive material.

Monitoring and auditing unstructured data to identify security vulnerabilities, risks, access rights, access patterns, and levels of protection is problematic because security information and event monitoring tools lack tight integration with identity management systems and file servers.

Secure reporting on which users and groups have been accessing and modifying data is insufficient. The reports should but often do not show chain of control of the file. And the reports themselves are often neither secure nor standardized.

Demonstrating chain of custody.

It is insufficient to merely find all your sensitive unstructured and control who can access it; you might also have to show the chain of custody. It can be a daunting problem to identify and demonstrate the chain of custody of a file or all the contents of a folder. Disk-based archiving for compliance and ediscovery can help, as can enterprise data management and data governance, but it can still be difficult to show the history of changes to access rights over time.

Each of these problems as well as some of the requirements and solutions that they entail are discussed in the sections below.

Consolidating Data Silos

Many large enterprises have inadequate centralized storage. Data silos — data stored on file servers that are segmented by protocol such as CIFS and NFS—resolutely block efforts to consolidate storage. Trying to bridge the gap between CIFS and NFS by using Samba to combine silos can be detrimental to network performance, complex to configure, and labor-intensive to maintain. The bottom line is that the data silos that accompany most heterogenous environments introduce layers of complexity, inefficiency, and insecurity.

"The result of too much heterogeneity is high management complexity and reduced efficiency," Andrew Reichman writes in his report, How Efficient Is Your Storage Environment? "Thinking about long-term economic effects of consolidation versus multisourcing makes good sense," he says.

One problem of tying different protocols to different security models is that the security descriptors on files can eventually clash, causing problems when the data is combined to be backed up on tape drives or archived in a central data warehouse.

Even some of the approaches to content management and archiving rely directly on file storage, further segmenting business data by content management system or archiving application. "It is always simpler to manage fewer vendors, products, solutions and technologies," Valdis Filks, a Gartner research analyst, writes in Best Practices for Turning a Storage Strategy Into Tactical Actions. "Therefore, the rationalization of the technologies used within your storage environment is a beneficial process."

Centralized storage goes beyond economies of scale, however. Large enterprises, whether government or business, are increasingly recognizing the short-term and long-term business value of the unstructured data that resides in files, driving them to "consolidate their file storage infrastructures to reduce overall costs and to improve the ability to protect and share all this data," Reichman writes in another report, File-Based Storage NAS Offerings And Other Approaches To Address An Avalanche Of Files.

But compatibility mismatches between platforms can still render efforts to consolidate storage problematic. "Storage teams should consider protocols against a backdrop of systems and platforms to determine whether they can interchange data and files without the need for specific and costly customizations and products. ... Unix and Linux formats may be interoperable with Windows formats, and CIFS files may need to be converted to work with Posix and NFS formats," Valdis Filks writes in his Gartner report on storage strategy.

In response to such problems, one trend of particular note is consolidation to a NAS system that supports both CIFS, which is the Windows file protocol, and NFS, which is the Unix file protocol. Doing so ensures interoperability, reduces complexity, and controls costs. And once unstructured data is consolidated in centralized storage, you can encrypt the files stored on the file server to enhance the security of your data at rest.

A multiprotocol, cross-platform file server can also help eliminate the need for storing unstructured data and business documents in collaboration tools and enterprise content management systems such as Documentum and SharePoint, frequently used by IT managers as a stopgap to help store sensitive files but just as frequently maligned by power users as unfit for storing business information that requires rapid access. And building applications to tap the data in such systems tends to result in proprietary lock and other problems. In addition, especially in large organizations, accessing centralized business data in enterprise content systems can be stymied by two constraints: obtaining permission to access it and overcoming platform-centric file protocols.

"Office productivity packages, much beloved for their flexibility, harbor a hidden curse: Users continually find clever ways to embed, paste, sort, and massage sensitive information into unstructured document form, where it continues to live forever, at rest and at risk....Enterprises should therefore provide speedy, centralized alternatives....Security should be a property of the solution — not the objective," Forrester analyst Andrew Jaquith writes in Own Nothing, Control Everything.

Consolidating your unstructured data into a common location of a streamlined storage system also provides the basis for being able to tap the mass of data later with big analytics or other applications for competitive advantage.

In addition, cross-platform file servers capable of handling multiple protocols from different originating platforms provide the flexibility to rapidly adapt to changing business needs. Supporting a wide variety of incompatible technologies and protocols is inefficient and expensive; consolidating storage to a cross-platform system simultaneously eliminates complexities, inefficiencies, management costs, and potential security problems.

But more importantly, cross-platform storage lays the foundation for many of the other technologies and processes that can help house and protect unstructured data. "Some data architectures lend themselves to distributed protection fairly easily. When data is stored in a central place, such as on a protected server (and is not allowed to be duplicated on a remote system), access control can also be managed centrally. Logs and monitors track access to the data and record which users and applications have viewed or altered it," Diana Kelley writes in a white paper from SecurityCurve called Addressing the Unstructured Data Protection Challenge.

Cross-platform, multiple-protocol file servers or NAS systems are the first step toward managing and securing the explosion of unstructured data with an extensible, cost-effective, flexible technology.

Locating, Classifying, and Structuring

Locating and understanding your unstructured data is an early step toward protecting it. "Knowing what you have and where it is correlates strongly with best-in-class performance in protecting and managing unstructured data," the Aberdeen Group writes in its report, Securing Unstructured Data: How Best-in-Class Companies Manage to Serve and Protect.

Many companies, however, do not spend enough resources to identify what unstructured data might be sensitive and where it is located — data that if compromised carries a significant risk.

In anything but the smallest business, setting out to discover and classify all your unstructured data would be a burdensome, resource-draining task. Automated discovery and classification tools can be slow and inaccurate; in addition, they are unlikely to be integrated well with your applications, data silos, and collaboration sites. Metadata standards such as Adobe XMP, Dublin Core, and Prism can help if your documents actually contain metadata, but many of them probably don't. In large enterprises, putting in place systems and processes for data governance and enterprise data management, while important, can take years — changes that can "get bogged down in lengthy implementation cycles and face adoption hurdles," Kelley says Addressing the Unstructured Data Protection Challenge.

Another common inhibitor to discovering and classifying unstructured data is unrealistic scoping. All your unstructured data need not be protected equally. Only sensitive data or data that you deem to be critical to your business requires the highest levels of protection.

There are several ways to solve this problem: Discriminately use software that discovers and categorizes files on a file server to determine whether they are potentially sensitive or not; empower departmental-level data stewards to tag folders and files by sensitivity level; and use compliance tagging to mark folders and files as sensitive in relation to a regulatory standard, such as Sarbanes-Oxley, PCI DSS, ITAR, HIPAA, and HITECH.

Once marked, either through automation, an originating application, or business processes, the sensitive data should be tracked on the file server — attempts to access, move, modify, or delete the sensitive data should lead to the steward or owner being alerted. Because the unstructured data can be tagged and tracked, a system of trusted viewing is put in place with security that is close to the data.

It’s something of a truism that to realize the full potential of your information assets while protecting sensitive files, unstructured data should be stored in repositories so that it can be easily found, accessed, and understood by users. In addition to tagging and tracking, the imperative for some form of structured storage of unstructured data is simple enough, old-fashioned and time-tested, stemming from the days of metal filing cabinets and paper folders: “The key to speedy electronic discovery” — which is especially important at large law firms and in corporate legal departments — “is having the documents properly filed and sorted in the first place, essentially to add some structure to their mass of unstructured data,” suggest a 2009 report from Grant Thornton, a consulting firm that specializes in corporate finance, risk management and information technology. There’s nothing like a well-organized file server governed by a security layer to provide users with fast, easy, secure access to documents.

Discovering and classifying unstructured data, however, is no longer sufficient. For sensitive data that falls under compliance regulations, you must also be able to show chain of custody — who had custody of and control over what files when. Without tracking ownership of and access to folders and files, it can be difficult to identify and prove the chain of custody — a failure to do so can result in fines and legal fees.

To track and prove chain of custody, your file servers storing your unstructured data must be tied to an access control system and a secure system of monitoring and reporting on file access, permissions changes, and file modifications over time.

Centralized Access Control and a Common Security Model

Immature, inconsistent security policies coupled with immature governance structures heighten risk. In large enterprises, the risks associated inconsistent security policies are made worse by failing to combine consistent policies with a single, centralized system for managing identities, authenticating users and groups, and controlling access to sensitive resources.

As a result, there are inadequate safeguards to shield unstructured data from prying eyes — even just read-only access. The consequences can be expensive: legal fees, expert fees, fines, compliance penalties, and declining revenues because of a tarnished reputation.

In the United States, for instance, the International Traffic in Arms Regulations, or ITAR, dictates that information and material pertaining to defense and military technologies may only be shared with U.S. persons unless exempted or specially authorized by the Department of State. Organizations can face huge legal fees and heavy fines if a foreign employee views ITAR-protected information.

The use of multiple identity management and access control systems — for example, using different directory services for different data silos — makes it difficult to implement uniform, consistent security policies that control access to unstructured data. Many businesses continue to use multiple identity and access control systems, such as NIS or LDAP for Unix and Active Directory for Windows computers. There is often one access control model for CIFS-based storage and another for NFS-based file servers.

A single system for managing identity, authentication, and access control provides a system-based approach for enforcing uniform, consistent security policies for computer users and groups, including managing their rights to access and modify unstructured data. But you also must be aware of having a negative impact.

As the Aberdeen Group points out, “a lack of clear ownership and a lack of consistent policies for unstructured data are particularly challenging” inhibitors to investing in protecting unstructured data. In his InformationWeek article called A Strategy to Protect Unstructured Data, Adam Ely adds that “protecting unstructured data is hard. To succeed, place controls close to the data and work outward, but be mindful of the impact of those controls on data owners and users....Efforts to secure unstructured data are just one facet of a larger layered security approach.”

But putting in place a consistent, uniform security model is difficult for good reason: There are legacy directory services, like NIS, to migrate off and there are barriers to extending a common access control system to file servers running on disparate operating systems, such as Linux, Solaris, Windows, and Mac OS X.

When those barriers can be overcome, there are the not-so-small matters of provisioning, managing, and deprovisioning user accounts across different platforms as well as implementing policies to manage users and groups who need privileged access for sensitive material stored in a variety of locations.

A requirement of PCI DSS, for instance, is to restrict access to cardholder data by business need-to-know. If a user account with access to sensitive unstructured data does not get deprovisioned from one of the access-control systems when the user leaves the company, it can violate compliance regulations.

Another problem is getting the levels of access just right: Flexible access with granular controls are key to disseminating information and tapping internal knowledge, both important indicators of success in a knowledge-based economy. Over-restricting access can constrict agility, undermine innovation, and stifle creativity. Protection must not unduly block the fast and furious flow of business.

Laying down a common security model that enforces consistent data-security policies is major step toward addressing who can access and modify sensitive data. A single identity management system can serve as the technical bridge by which consistent policies can be implemented to improve access controls. It provides a system-based approach for implementing uniform, consistent security policies for users and departments, including managing their access rights for unstructured data.

Because a file server represents a single point of control for a security policy, using a single identity management and access control system makes for a sound security framework: Changes to access rights for users and groups typically take effect immediately.

Identifying Who can Access What

Unstructured data that is distributed across the enterprise, segmented into storage silos, and controlled by disparate access control systems compound the problem of determining the users and groups who can access sensitive material. Consolidating storage of sensitive unstructured data on to file servers governed by a common security model and access control system can help identify who has access to what data, providing a framework to better regulate access at a granular level.

Identifying user and group access to sensitive assets spans several of the requirements for securing unstructured data discussed in this white paper. Consolidating data silos into a cross-platform file server or NAS system gives you heightened visibility into where your data resides, which in turn helps you more easily locate, classify, and categorize your data as well as control access to it with a common framework. The combination not only gives you the ability to identify which users and groups can access what data, but enables monitoring and reporting frameworks to be put in place so that you can demonstrate who can access what and monitor users who attempt to access or modify the files.

Monitoring Security Events

Why is monitoring so important? In a word: Intelligence. Monitoring and auditing unstructured data empowers you to identify security vulnerabilities, risks, overly permissive access rights, patterns of access, levels of protection, and changes to files and folders marked sensitive.

On the file server, a component that records information about moving, copying, reading, modifying, or deleting directories or files can help fulfill compliance requirements. PCI DSS Requirement 10, for instance, requires among other things that you record at least the following audit trail entries for all system components for each event containing sensitive credit card information:

- User identification
- Type of event
- Date and time
- Success or failure indication
- Origination of event
- Identity or name of affected data, system component, or resource

With the requirements of the PCI security standard, the stakes are particularly high. In addition to bad press, security breaches can lead to fines that run up to $500,000 or more.

In large organizations, administrators and end users are frequently unaware of regulatory requirements for sensitive data. Unless automated systems are put in place to force adherence and to monitor for lapses, they will inadvertently subvert those requirements.

Monitoring patterns of access and activity provides context-specific insights and forensics that you can act on to mitigate risk, comply with regulations, make access control decisions, and ward off potential security breeches in real time. When monitoring detects certain defined security events, alerts should prompt administrators or users to take some form of action, such as sending an automated email to inform the user about a potential violation of policy. In short, monitoring can help control the misuse of unstructured data.

To ensure that events do not consume too much network traffic or bog down systems, monitoring ultimately should take place as part of the file server. In addition, as Forrester Research analyst Andras Cser has pointed out in an Active Directory Q&A, administrators who integrate security monitoring deeply into their network infrastructure will benefit from scale efficiencies and cost reductions.

Monitoring is more than just a means to comply with regulations, however. It produces large amounts of data about patterns of access and activity — data that becomes an input to an important use that progressive organizations are exploiting to drive innovation, create value, and become more efficient: Analytics.

An analytics system can, in turn, use data about past access patterns and file activities to hypothesize about future patterns of data storage. The inferences of an analytics system can help identify files that might contain sensitive data and need to be flagged for inspection or tracking.

By monitoring the union of identity, role, entitlement, and file activity, you can not only optimize your security for sensitive assets but also satisfy the demands of regulatory compliance and identify potential breaches in real time.

Reporting and Auditing

Reporting and auditing can help identify security vulnerabilities, mitigate security risks, inspect access rights, track patterns of access and change, and double-check levels of protection -- all of which can come into to play to help prove compliance with regulations such as PCI, SOX, and ITAR.

Some large companies annually audit the data-handling controls of the smaller companies from which they purchase and embed material or components, Forrester analyst Andrew Jaquith writes in Own Nothing, Control Everything, adding that some federal government agencies follow an assurance standard such as ISO 17799 and version of SAS 70, and may reserve the right to audit their contractors. Having your own auditing system in place to monitor and inspect unstructured data helps reduce the possibility of surprises later.

But reporting and auditing is not just about proving compliance, it's also about cutting costs. According to the Aberdeen Group, "The greatest financial gains for best-in-class organization will come from automating the enforcement of policies whenever reasonable, standardizing audit, analysis, and reporting, and driving continuous improvements by finding and eliminating root causes for exceptions, security events, and audit deficiencies."

Yet many organizations lack reports tied to security information and event monitoring (SIEM) tools. Integrating SIEM tools that isolate file events with a reporting system lets you generate reports to show which users changed what files over time. In addition, few organizations have integrated their reporting and auditing tools with their identity management and access control systems. Even fewer organizations move beyond reports to use dashboards to monitor correlated file server events in real-time.

More importantly, linking the reporting system to SIEM tools as well as the identity management and access control systems empowers you to show who owned and modified sensitive files over time — endowing you with a framework to show chain of custody.

For many regulations, the reporting system itself as well as the reports that are generated must be secured with access control, typically as part of a policy that addresses information security for all personnel. A reporting system that is integrated with the identity management system allows you to do so, effectively enabling to show change logs and chain of custody not only for sensitive data but also for the reports themselves.

Standardizing and automating reports at the confluence of storage, identity, and access radically improves visibility to possible data breeches, security threats, compliance failures, and other transgressions that expose your organization to risk.

Conclusion

The companies that performed at the top of Aberdeen Group's survey discovered and classified their unstructured data and then established consistent policies to control its access, distribution, and use. The policies were then enforced through training and "standardizing the use of data protection technologies." Companies who put in place systems and policies to protect unstructured data, the Aberdeen Group report says, reaped rewards by reducing the costs associated with data management, data loss, security breeches, compliance shortcomings, audit deficiencies, and help desk tickets.

A number of requirements — centralized storage, identity management and access control, object tracking, event monitoring, and secure auditing and reporting — come together to provide the foundation for securing sensitive unstructured data. Consolidating data silos into cross-platform, multiprotocol file servers or NAS systems reduces costs while easing the implementation of a common security model and making it easier to locate and classify sensitive information.

A single system for managing identities and authenticating users is the basis for a common security model to enforce consistent data-security policies to control who can access and modify sensitive data.

Monitoring and auditing serve as a layer of backstop security to identify security vulnerabilities, mitigate risks, inspect access rights, track access patterns, and double-check levels of protection. Meantime, secure reporting helps demonstrate regulatory compliance and show chain of control over sensitive data that you chosen to track.

By uniting identity, security, and storage, your sensitive unstructured data is free to be used by those who need it but secured from the prying eyes of those who do not.

FOR MORE INFORMATION

For more information on Likewise, visit the web site at http://www.likewise.com. To contact the sales team, call (800) 378-1330 or e-mail info@likewise.com.

ABOUT LIKEWISE STORAGE SERVICES

Likewise Storage Services provides industrial-strength client and server SMB/CIFS support so Microsoft Windows clients can access folders and files on Linux, Unix and Mac computers. At the same time, the Likewise Storage Services FUSE module enables a Linux computer to access folders and files. Administrators can remotely manage Likewise Storage Services File Server on Linux and Unix machines by using popular Microsoft Windows tools. The Likewise SMB File Server supports both SMB1 and SMB2 and can be integrated with proprietary technologies.

ABOUT LIKEWISE

Likewise makes an integrated software platform for identity, security and storage used by market-leading storage vendors including HP and EMC. Likewise enables organizations to provide both access to and control of their data across mixed network environments. More information is available at the company’s web site,www.likewise.com.

The information contained in this document represents the current view of Likewise Software on the issues discussed as of the date of publication. Because Likewise Software must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Likewise, and Likewise Software cannot guarantee the accuracy of any information presented after the date of publication. The contents herein are subject to change without notice.

These documents are for informational purposes only. LIKEWISE SOFTWARE MAKES NO WARRANTIES, EXPRESS OR IMPLIED.

Likewise Software
15395 SE 30th Place, Suite #140
Bellevue, WA 98007
USA