Data Leakage and Prevention

Module II

Dr. P. Vijayakumar

Introduction Data Leakage

- Data leakage is defined as the **accidental or unintentional distribution of private or sensitive data to an unauthorized entity.**

- Sensitive data in companies and organizations include Intellectual Property (IP), Financial Information, Patient Information, Personal Credit Card Data, and other information depending on the business and the industry.

- Data leakage poses a **serious issue** for companies as the number of incidents and the cost to those experiencing them continue to increase.

Introduction Data Leakage

- Data leakage is enhanced by the fact that transmitted data (both inbound and outbound), including emails, instant messaging, website forms and file transfers among others, are largely unregulated and unmonitored on their way to their destinations.

- Sensitive data are shared among various stakeholders such as employees working from outside the organization’s premises (e.g. on laptops), business partners and customers.

- This increases the risk that confidential information will fall into unauthorized hands.

Introduction to Data Leakage

- The potential damage and adverse consequences of a data leakage incident can be classified into two categories:

- Direct losses
  - Direct losses refer to tangible damage that is easy to measure or to estimate quantitatively.

- Indirect losses.
  - Harder to quantify and have a much broader impact in terms of cost, place and time

Direct losses

- Direct losses include violations of regulations (such as those protecting customer privacy) resulting in
  - fines;
  - settlements or customer compensation fees;
  - litigation involving lawsuits;
  - loss of future sales;
  - costs of investigation and
  - remedial or restoration fees.

Indirect losses

- Indirect losses include
  - reduced share price as a result of negative publicity;
  - damage to a company's goodwill and reputation;
  - customer abandonment;
  - and exposure of intellectual property (business plans, code, financial reports and meeting agendas) to competitors.

Data Leakage Prevention (DLP)

- Enterprises use Data Leakage Prevention (DLP) technology as one component in a comprehensive plan for the handling and transmission of sensitive data.

- The technological means employed for enhancing DLP can be divided into the following categories:

  - Standard security measures
  - Advanced/ intelligent security measures
  - Access control and encryption
  - Designated DLP systems

Standard Security Measures

- Standard security measures are used by many organizations and include common mechanisms such as
  - firewalls,
  - intrusion detection systems (IDSs) and
  - antivirus software
- It can provide protection against both
  - outsider attacks (e.g. a firewall which limits access to the internal network and an intrusion detection system which detects attempted intrusions) and
  - inside attacks (e.g. antivirus scans to detect a Trojan horse that may be installed on a PC to send confidential information).

Data Leakage Prevention (DLP) System

- Another example is the use of thin clients which operate in a client-server architecture, with **no personal or sensitive data** stored on a client's computer.

- **Policies and training** for improving the awareness of employees and partners provide additional standard security measures.

Advanced or intelligent security measures include

- machine learning and temporal reasoning algorithms for detecting abnormal access to data (i.e. databases or information retrieval systems),
- activity based verification (e.g. based on keystrokes and mouse patterns),
- detection of abnormal email exchange patterns, and
- applying the honeypot concept for detecting malicious insiders.

Data Leakage Prevention (DLP) System

- *Device control*, *access control and encryption* are used to prevent access by an unauthorized user.

- These are the simplest measures that can be taken to protect large amounts of personal data against malicious outsider and insider attacks.

Designated DLP solutions

- *Designated DLP solutions* are intended to detect and prevent attempts to copy or send sensitive data, intentionally or unintentionally, without authorization, mainly by personnel who are authorized to access the sensitive information.

- A major capability of such solutions is an ability to classify content as sensitive.

- Designated DLP solutions are typically implemented using mechanisms such as exact data matching, structured data fingerprinting, statistical methods (e.g. machine learning), rule and regular expression matching, published lexicons, conceptual definitions and keywords.

Data Leakage Prevention (DLP) solutions

- A designated data leakage prevention solution is defined as a system that is designed to detect and prevent the unauthorized access, use or transmission of confidential information.

  - Information Leak Prevention (ILP),
  - Data Leak/ Loss Prevention (DLP),
  - Outbound Content Compliance,
  - Content Monitoring and Filtering,
  - Content Monitoring and Protection (CMP) or Extrusion Prevention.

Enterprise data generally exists in the following three major states:

- **Data at rest:** it resides in files systems, distributed desktops and large centralized data stores, databases or other storage centers.

- **Data at the endpoint or in use:** it resides at network endpoints such as laptops; USB devices; external drives; CD/ DVDs; archived tapes; MP3 players; iPhones or other highly mobile devices.

- **Data in motion:** it moves through the network to the outside world via email, instant messaging, peer-to-peer (P2P), FTP or other communication mechanisms.

Organizational Data Classification, Location and Pathways

- Enterprises are often unaware of all of the types and locations of information they possess.

- It is important, prior to purchasing a DLP solution, to identify and classify sensitive data types and their flow from system to system and to users.

- This process should yield a data taxonomy or classification system that will be leveraged by various DLP modules as they scan for and take action on information that falls into the various classifications within the taxonomy.

Organizational Data Classification, Location and Pathways

- Analysis of critical business processes should yield the required information.
- Classifications can include categories such as private customer or employee data, financial data and intellectual property.
- Once the data have been identified and classified appropriately, further analysis of processes should facilitate the location of primary data stores and key data pathways.
- Frequently multiple copies and variations of the same data are scattered across the enterprise on servers, individual workstations, tape and other media.

Organizational Data Classification, Location and Pathways

- Copies are frequently made to facilitate application testing without first cleansing the data of sensitive content.

- Having a good idea of the data classifications and location of the primary data stores proves helpful in both the selection and placement of the DLP solution.

- Once the DLP solution is in place, it can assist in locating additional data locations and pathways.

- It is also important to understand the enterprise's data life cycle.

Organizational Data Classification, Location and Pathways

- Understanding the life cycle from point of origin through processing, maintenance, storage and disposal will help uncover further data repositories and transmission paths.

- Additional information should be collected by conducting an inventory of all data egress points since not all business processes are documented and not all data movement is a result of an established process.

- Analysis of firewall and router rule sets can aid these efforts.

DLP features vs. DLP solutions

- The DLP market is also split between DLP as a feature and DLP as a solution.
- A number of products, particularly email security solutions, provide basic DLP functions, but aren't complete DLP solutions.
- The difference is:
- A DLP product includes centralized management, policy creation and enforcement workflow dedicated to the monitoring and protection of content and data.

- The user interface and functionality are dedicated to solving the business and technical problems of protecting content through content awareness.

- DLP features include some of the detection and enforcement capabilities of DLP products, but are not dedicated to the task of protecting content and data.

Content Awareness

- **Content vs. Context**
  - One of the defining characteristics of DLP solutions is their content awareness.
  - This is the ability of products to analyse deep content using a variety of techniques, and is very different from analyzing context.
  - It's easiest to think of content as a letter and context as the envelope and environment around it.
- Context includes things like source; destination; size; recipients; sender; header information; metadata; time; format and anything else short of the content of the letter itself.
- A more advanced version of contextual analysis is business context analysis, which involves deeper analysis of the content, its environment at the time of analysis and the use of the content at that time.

Content Awareness

- Content awareness involves peering inside containers and analysing the content itself.

- The advantage of content awareness is that while we use context, we're not restricted by it.

- If I want to protect a piece of sensitive data, I would want to protect it everywhere and not just in obviously sensitive containers.
- I'm protecting the data, not the envelope, so it makes a lot more sense to open the letter, read it, and decide how to treat it.

- This is more difficult and time consuming than basic contextual analysis and is the defining characteristic of DLP solutions.

Content Analysis

- The first step in content analysis is capturing the envelope and opening it.

- The engine then needs to parse the context (we'll need that for the analysis) and dig into it.

- This is easy for a plain text email, but when you want to look inside binary files, it gets a little more complicated.

- All DLP solutions solve this using file cracking. File cracking is the technology used to read and understand the file, even if the content is buried multiple levels down.

Content Analysis

- For example, it's not unusual for the cracker to read an Excel spreadsheet embedded in a Word file that's zipped.

- The product needs to unzip the file, read the Word doc, analyse it, find the Excel data, read it and analyse it.

- Other situations get far more complex, like a .pdf embedded in a CAD file.

- Many of the products in the market today support around 300 file types, embedded content, multiple languages, double byte character sets for Asian languages, and pulling plain text from unidentified file types.

Content Analysis

- Quite a few use the autonomy or verity content engines to help with file cracking, but all the serious tools have quite a bit of proprietary capability, in addition to the embedded content engine.

- Some tools support analysis of encrypted data if enterprise encryption is used with recovery keys, and most tools can identify standard encryption and use that as a contextual rule to block/ quarantine content.

Content Analysis Techniques

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

1. Rule based/ Regular expressions:

2._Database fingerprinting:

3._Exact file matching

4._Partial document matching

5._Statistical analysis:

6._Conceptual/ Lexicon

7._Categories:

Rule based/ Regular expressions

- This is the most common analysis technique available in both DLP products and other tools with DLP features.

- It analyses the **content for specific rules**, such as 16 digit numbers that meet credit card checksum requirements, medical billing codes or other textual analyses.

- Most DLP solutions enhance **basic regular expressions with their own additional analysis rules** (e.g. a name in proximity to an address near a credit card number).

- Its advantages are: 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.

Rule based/ Regular expressions

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

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

Database fingerprinting

- Sometimes called Exact Data Matching – this technique takes either **a database dump or live data (via ODBC connection) from a database and only looks for exact matches**.

- For example, you could generate a policy to look only for credit card numbers in your customer base, thus ignoring your own employees buying online.

- More advanced tools look for combinations of information, such as the magic combination of first name or initial with last name, credit card or social security number that triggers a disclosure.

- Make sure you understand the performance and security implications of nightly extracts vs. live database connections.

- **Its advantages** are: structured data from databases.

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

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

Exact file matching

- With this technique you take a hash of a file and monitor for any files that match that exact fingerprint.

- Some consider this to be a contextual analysis technique since the file contents themselves are not analysed.

- **Its advantages are**: media files and other binaries where textual analysis isn't necessarily possible.

- **Strength**s: works on any file type, low false positives with a large enough hash value (effectively none).

- **Weaknesse**s: trivial to evade. Worthless for content that's edited, such as standard office documents and edited media files.

Partial document matching

- This technique looks for a **complete or partial match on protected content.**

- Thus you could build a policy to protect a sensitive document, and the DLP solution will look for either the complete text of the document, or even excerpts as small as a few sentences.

- For example, you could load up a business plan for a new product and the DLP solution would alert if an employee pasted a single paragraph into an Instant Message.

- Most solutions are based on a technique known as cyclical hashing, where you take a hash of a portion of the content, offset a predetermined number of characters, then take another hash, and keep going until the document is completely loaded as a series of overlapping hash values.

Partial document matching

- Its advantages are: protecting sensitive documents or similar content with text such as CAD files (with text labels) and source code. Unstructured content that's known to be sensitive.

- Strengths: ability to protect unstructured data. Generally low false positives (some vendors will say zero false positives, but any common sentence/ text in a protected document can trigger alerts).

- Doesn't rely on complete matching of large documents. It can find policy violations on even a partial match.

- Weaknesses: performance limitations on the total volume of content that can be protected. Common phrases/ verbiage in a protected document may trigger false positives. Must know exactly which documents you want to protect.

Statistical analysis

- Use of machine learning, Bayesian analysis and other statistical techniques to analyse a corpus of content and find policy violations in content that resembles the protected content.

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

- Some techniques are very similar to those used to block spam.

- Its advantages are: unstructured content where a deterministic technique, like partial document matching would be ineffective

Statistical analysis

- For example, a repository of engineering plans that's impractical to load for partial document matching due to high volatility or massive volume.

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

- Weaknesses: prone to false positives and false negatives. Requires a large corpus of source content – the bigger, the better.

Conceptual/ Lexicon

- This technique uses a combination of dictionaries, rules and other analyses to protect nebulous content that resembles an "idea".

- It's easier to give an example — a policy that alerts on traffic that resembles insider trading, which uses key phrases, word counts and positions to find violations.

- Other examples are sexual harassment, running a private business from a work account and job hunting.

Conceptual/ Lexicon

- Its advantages are: completely unstructured ideas that defy simple categorization based on matching known documents, databases or other registered sources.

- Strengths: not all corporate policies or content can be described using specific examples.

- Conceptual analysis can find closely defined policy violations other techniques can't even think of monitoring for.

- Weaknesses: in most cases, these are not user-definable and the rule sets must be built by the DLP vendor with significant effort, which costs more. This technique is very prone to false positives and negatives because of the flexible nature of the rules.

Categories

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

- Its advantages are: anything that neatly fits a provided category. Typically, easy to describe content related to privacy, regulations or industry specific guidelines.

Categories

- Strengths: extremely simple to configure. Saves significant policy generation time.
- Category policies can form the basis for more advanced, enterprise specific policies.
- For many organizations, categories can meet a large percentage of their data protection needs.
- Weaknesses: one size fits all might not work. Only good for easily categorized rules and content.

Categories

- These seven techniques form the basis for most of the DLP products on the market. Not all products include all techniques, and there can be significant differences between implementations.

- Most products can also chain techniques — building complex policies from combinations of content and contextual analysis techniques.

Data Protection

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

- **Data at Rest** includes scanning of storage and other content repositories to identify where sensitive content is located.

- We call this content discovery. For example, you can use a DLP product to scan your servers and identify documents with credit card numbers.

- If the server isn't authorized for that kind of data, the file can be encrypted or removed or a warning sent to the file owner.

Data in Motion

- **Data in Motion** is sniffing of traffic on the network (passively or inline via proxy) to identify content being sent across specific communications channels.

- For example, this includes sniffing emails, instant messages and web traffic for snippets of sensitive source code.

- In motion, tools can often block based on central policies depending on the type of traffic.

Data in Use

- **Data in Use** is typically addressed by endpoint solutions that monitor data as the user interacts with it.

- For example, they can identify when you attempt to transfer a sensitive document to a USB drive and block it (as opposed to blocking use of the USB drive entirely).

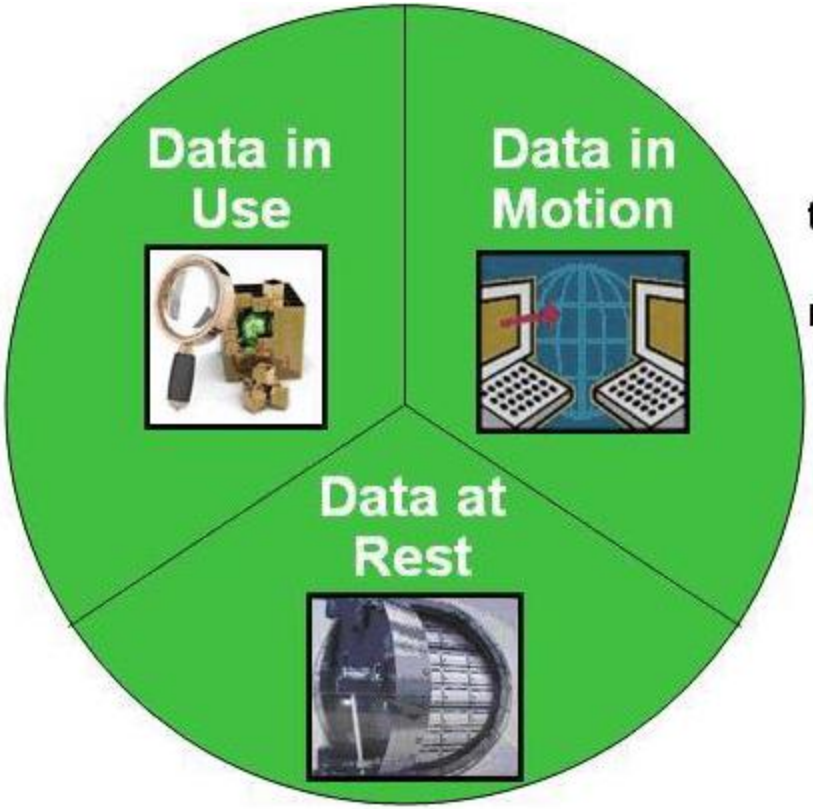
- Data in use tools can also detect things like copy and paste or use of sensitive data in an unapproved application (such as someone attempting to encrypt data to sneak it past the sensors).

- Many organizations first enter the world of DLP with network based products that provide broad protection for managed and unmanaged systems.

- It’s typically easier to start a deployment with network products to gain broad coverage quickly.

- Early products limited themselves to basic monitoring and alerting, but all current products include advanced capabilities to integrate with existing network infrastructure and provide protective, not just detective controls.

Data in Use:
Active data under constant change stored physically in databases, data warehouses, spreadsheets etc.



Data in Motion:
Data that is traversing a network or temporarily residing in computer memory to be read or updated.

Data at Rest:
Inactive data stored physically in databases, data warehouses, spreadsheets, archives, tapes, off-site backups etc.

Data In Motion

Network Monitor

- At the heart of most DLP solutions lies a passive network monitor. The network monitoring component is typically deployed at or near the gateway on a SPAN port (or a similar tap).
- It performs full packet capture, session reconstruction and content analysis in real time.
- Performance is more complex and subtle than vendors normally discuss.
- First, on the client expectation side, most clients claim they need full gigabit ethernet performance, but that level of performance is unnecessary except in very unusual circumstances since few organizations are really running that high a level of communications traffic.

Data In Motion

- DLP is a tool to monitor employee communications, not web application traffic. Realistically, we find that small enterprises normally run under 50 MByte/s of relevant traffic, medium enterprises run closer to 50-200 MB/s and large enterprises around 300 MB/s (maybe as high as 500 in a few cases).,

- Not every product runs full packet capture because of the content analysis overhead. You might have to choose between pre-filtering (and thus missing non-standard traffic) or buying more boxes and load balancing.

- Also, some products lock monitoring into pre-defined port and protocol combinations, rather than using service/ channel identification based on packet content.

Data In Motion

- Even if full application channel identification is included, you want to make sure it's enabled otherwise you might miss nonstandard communications such as connecting over an unusual port.

- Most of the network monitors are dedicated general purpose server hardware with DLP software installed.

- A few vendors deploy true specialized appliances.

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

Email Integration

- The next major component is email integration. Since email is stored and forwarded, you can gain a lot of capabilities, including quarantine, encryption integration and filtering without the same hurdles to avoid blocking synchronous traffic.

- Most products embed an MTA (Mail Transport Agent) into the product, allowing you to just add it as another hop in the email chain.

- Quite a few also integrate with some of the major existing MTAs/ email security solutions directly for better performance.

Email Integration

- One weakness of this approach is it doesn't give you access to internal email.
- If you're on an exchange server, internal messages never make it through the external MTA since there's no reason to send that traffic out.
- To monitor internal mail, you'll need direct Exchange/ Lotus integration, which is surprisingly rare in the market.
- Full integration is different from just scanning logs/ libraries after the fact, which is what some companies call internal mail support.
- Good email integration is absolutely critical if you ever want to do any filtering, as opposed to just monitoring.

Filtering/ Blocking and Proxy Integration

- Nearly anyone deploying a DLP solution will eventually want to start blocking traffic.

- There's only so long you can take watching all your sensitive data running to the nether regions of the Internet before you start taking some action.

- Blocking isn't the easiest thing in the world, especially since we're trying to allow good traffic.

- Block only bad traffic, and make the decision using real-time content analysis.

Filtering/ Blocking and Proxy Integration

- Email, as we mentioned, is fairly straightforward to filter.

- It's not quite real time and is ‘proxied’ by its very nature.
- Adding one more analysis hop is a manageable problem in even the most complex environments.
- Outside of email, most of our communications traffic is synchronous.
- Everything runs in real time. Thus if we want to filter it we either need to bridge the traffic, proxy it or poison it from the outside.

Bridge

- With a bridge, we just have a system with two network cards which performs content analysis in the middle.
- If we see something bad, the bridge breaks the connection for that session.
- Bridging isn't the best approach for DLP since it might not stop all the bad traffic before it leaks out.
- It's like sitting in a doorway watching everything go past with a magnifying glass.
- By the time you get enough traffic to make an intelligent decision, you may have missed the really good stuff.
- Very few products take this approach although it does have the advantage of being protocol agnostic.

Proxy

- In simplified terms, a proxy is protocol/ application specific and queues up traffic before passing it on, allowing for deeper analysis.
- We see gateway proxies mostly for HTTP, FTP and IM protocols.
- Few DLP solutions include their own proxies.
- They tend to integrate with existing gateway/ proxy vendors since most customers prefer integration with these existing tools.
- Integration for web gateways is typically through the iCAP protocol, allowing the proxy to grab the traffic, send it to the DLP product for analysis and cut communication, if there's a violation.

Proxy

- This means you don't have to add another piece of hardware in front of your network traffic, and the DLP vendors can avoid the difficulties of building dedicated network hardware for inline analysis.

- If the gateway includes a reverse SSL proxy you can also sniff SSL connections.

- You will need to make changes on your endpoints to deal with all the certificate alerts, but you can now peer into encrypted traffic.

- For Instant Messaging, you'll need an IM proxy and a DLP product that specifically supports whatever IM protocol you're using.

TCP Poisoning

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

Internal Networks

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

Distributed and Hierarchical Deployments

- All medium to large enterprises and many smaller organizations have multiple locations and web gateways.

- A DLP solution should support multiple monitoring points, including a mix of passive network monitoring, proxy points, email servers and remote locations.

- While processing/ analysis can be offloaded to remote enforcement points, they should send all events back to a central management server for workflow, reporting, investigations and archiving.

- Remote offices are usually easy to support since you can just push policies down and reporting back, but not every product has this capability.

Distributed and Hierarchical Deployments

- The more advanced products support hierarchical deployments for organizations that want to manage DLP differently in multiple geographic locations or by business unit.

- International companies often need this to meet legal monitoring requirements which vary by country.

- Hierarchical management supports coordinated local policies and enforcement in different regions, running on their own management servers and communicating back to a central management server.

Distributed and Hierarchical Deployments

- Early products only supported one management server but now we have options to deal with these distributed situations with a mix of corporate/ regional/ business unit policies, reporting and workflow.

Data At Rest

- While catching leaks on the network is fairly powerful, it's only one small part of the problem.

- Many customers are finding that it's just as valuable, if not more valuable, to figure out where all that data is stored in the first place.

- We call this content discovery. Enterprise search tools might be able to help with this, but they really aren't tuned well for this specific problem.

- Enterprise data classification tools can also help, but based on discussions with a number of clients, they don't seem to work well for finding specific policy violations.

Data At Rest

- Thus we see many clients opting to use the content discovery features of their DLP products.

- The biggest advantage of content discovery in a DLP tool is that it allows you to take a single policy, and apply it across data no matter where it's stored, how it's shared, or how it's used.

- For example, you can define a policy that requires credit card numbers to only be emailed when encrypted, never be shared via HTTP or HTTPS, only be stored on approved servers and only be stored on workstations/ laptops by employees on the accounting team.
- All of this can be specified in a single policy on the DLP management server.