

ERNW Newsletter 29 / November 2009

Dear Partners and Colleagues,

Welcome to the ERNW Newsletter no. 29 covering the topic:

Data Leakage Prevention — A Practical Evaluation

Version 1.0 from 19th of november 2009

Author: Matthias Luft, mluft@ernw.de

Abstract

This newsletter illustrates the new technology Data Leakage Prevention (DLP). After some basic definitions and theoretical explanations, the evaluation of two exemplary DLP suites is described in detail. This examination is based on several requirements and derived test cases, which cover most aspects of DLP and can also serve as a framework for further examinations.

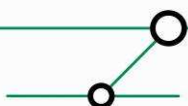
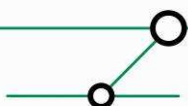


TABLE OF CONTENTS

1	INTRODUCTION.....	3
1.1	Some words on leakage.....	3
1.2	Outline.....	3
2	DATA LEAKAGE PREVENTION.....	3
2.1	Identify: How To Find Valuable Content.....	4
2.2	Monitor: Controlling Every Channel.....	4
2.3	React: Handling Policy Breaches.....	5
3	TESTING METHODOLOGY.....	5
3.1	Test Cases.....	5
4	EVALUATION.....	6
4.1	Concrete Test Cases.....	6
4.2	Websense Data Security Suite.....	7
4.2.1	MIME Types Not Examined In Depth.....	7
4.2.2	Missing Analysis of Meta Data.....	7
4.2.3	Freeze During Analysis.....	7
4.2.4	Insufficient Encryption.....	8
4.3	McAfee Host Data Loss Prevention.....	9
4.3.1	MIME Types Not Examined In Depth.....	9
4.3.2	Missing Analysis of Meta Data.....	9
4.3.3	Insufficient React Behavior.....	9
4.3.4	Insufficient Monitoring of USB Hard Drives.....	9
4.4	Summary.....	9
5	GENERAL PROBLEMS AND CONCLUSION.....	9



1 INTRODUCTION

DLP is the general term for a new approach to avoid data breaches. To achieve this aim, all currently available implementations perform analysis of intercepted data. This analysis is based on defined policies which describe valuable data. There are different possibilities to both define these content policies and to intercept data to make the analysis possible.

Since the DLP market is still adolescent, new DLP suites are rising every year. In the context of an evaluation for a customer who wants to implement such content analysis software, we examined two of these solutions. The following newsletter describes the results of this examination.

1.1 Some words on leakage

When concerning with DLP, it is important to define the term Data Leakage. First of all, the occurrence of data leakage is independent from the impact on the confidentiality of information. There are examples when encrypted data was lost by companies by incident which leads to serious business impact even there was no information leaked. Furthermore, leakage can occur in different situations. An attacker which carries away sensible data leads to leakage as well as an employee who wants to get its job done. So the following, very general, definition of data leakage can be applied:

Data Leakage is – from the owner's point of view – unintentional loss of confidentiality for any kind of data.

1.2 Outline

The next chapters will explain what DLP is and how it is to protect against data breaches. Chapter 2 describes general concepts how data is intercepted and analyzed in the context of DLP. This theoretical groundwork is necessary to understand why the used test cases (Chapter 3) are adequate and cover the important capabilities of a DLP suite. These developed test cases are used to evaluate the two exemplary DLP solutions. After some basic information on the solutions, the gathered results are listed in Chapter 4.

2 DATA LEAKAGE PREVENTION

This chapter explains current approaches that are implemented in DLP solutions. The following definition of DLP is a good source to point out all important aspects:

Products that, based on central policies, identify, monitor, and protect data at rest, in motion, and in use, through deep content analysis¹

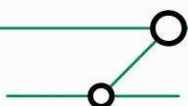
Based on this definition, it is possible to derive the three main capabilities of DLP solutions:

- Identify
- Monitor
- React

Each of these steps in leakage prevention has to deal with the mentioned requirements to handle data

- at rest,
- in motion,
- and in use.

¹ Rich Mogull. *Understanding and selecting a data loss prevention solution. Technical report, SANS Institute, 2007.*



Thus the remainder of this chapter explains in depth how the different challenges are handled in each situation. There are several products which address single requirements, like to scan for certain content. In contrast, complete DLP solutions should cover all of these tasks that can also be fulfilled by single programs. This approach allows the central management of all components and thus the sharing of results for further steps. Nevertheless, the remainder of this chapter explains in depth how the different work stages are handled by the complete DLP suites. This allows a more fine grained analysis of the single capabilities and is necessary to identify possible points of failure.

2.1 Identify: How To Find Valuable Content

If sensitive data should be protected, every kind of control mechanisms needs to know how the valuable data looks like. So in a first step, methods of defining data and scanning for it are needed. It is not practicable to insert every piece of information that is worthy of protection into, for example, a database. A central management is needed since the policies must be consistent and manageable. This is not provided when all policies are spread through different places or tools. It is also necessary to provide generic methods to define data both as general and as special as needed. The following technologies provide capabilities to describe data in various ways. It is important for a DLP solutions to implement as much of these technologies as possible since only a reasonable combination of them leads to satisfying, real-world detection rates.

Rule-Based

Regular Expressions are the most common technique for defining data via an abstract pattern. At the same time, this is the biggest constraint since this approach produces a high rate of false positives due to the limited scope and missing context awareness.

Database Fingerprinting

If it is possible to identify a database that holds lot of sensitive data, this database can be read to perform exact file matching.

Exact File Matching

Like the extraction out of a database, existing amounts of data, e.g., on a file server can be hashed and indexed. Using these footprints, matching can be performed on any kind of file types with a low rate of false positives.

Partial Document Matching

This technique processes complete documents to be able to match particular appearances in other documents. So every part of sensible documents can be matched even if they are only partially included in other documents.

Statistical Analysis

Some DLP solutions use modern machine learning techniques either by processing an amount of training data or by learning continuously at work. Exactly as is the case with learning spam filters, this approach will lead to false positives and false negatives.

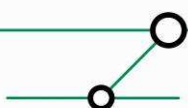
Conceptual

If it is possible to define a dictionary of sensible data, the DLP solution can judge content based on the contained words. So unstructured but sensible or unwanted data can be assessed and scored.

All of these approaches analyze content of data. The solution must thus have the ability to *understand* lots of different file types like Word documents, zip archives or images. This deep content analysis allows the further processing using the methods mentioned above.

2.2 Monitor: Controlling Every Channel

In a second step, data must be accessible to allow the application to apply any kind of control. Since the last section was focused on the content analysis, this section regards the context of data. The context of data is again related to the different states of data: In motion, at rest, and in use.



In Motion

Data in motion basically means every data that is transferred over the network. So there are several monitoring points to intercept traffic. This includes both integration in web proxies or mail servers and passive network monitoring using a special network port or similar controls.

At Rest

The scanning of data at rest is one of the most important use cases for a DLP solution. So an organization can recognize where its sensitive data is distributed all about. Most of the scanning can be done using network sharing methods, but in some cases as the assessment of endpoint systems or application servers, a software agent is needed to access all data.

In Use

Since every user behavior is a different use case for leakage prevention, it is not possible to remotely monitor data in use. To control every action a user may take, an endpoint agent is needed. This agent hooks up important operating system functions to recognize all actions, like copying to the clipboard, a user takes.

2.3 React: Handling Policy Breaches

There are several controls to handle the different kinds of detected data leakage. Depending on the state of data, it is necessary to have an appropriate reaction policy. It is not appropriate to delete sensible documents which are found on a public file server – this would lead to a radical decrease of the availability of data even if it would avoid any leakage. There must exist fine grained possibilities to determine what controls should be applied. In the case of the file server, the file should get moved to a secure place leaving a *moved to* message. Encryption is also a way to secure discovered data: Providing a challenge code the password for decryption can be requested. The data then has to be moved to another place to avoid further policy breaches.

3 TESTING METHODOLOGY

When evaluating any piece of software, the test cases derive from its application and specification. The test scenarios of a DLP endpoint solution therefore must cover all use cases which represent typical user behavior that could affect the confidentiality of data. In doing so, both intentional and unintentional leakage of data must be handled properly. Intentional data leakage includes firstly malicious activities, but also an employee who is restricted by the DLP solution and wants to get his work done, e.g., by sending an email containing important information to a colleague.

These use cases lead to concrete technical checks which are summarized in the next section. The classification derives from the different stages of a DLP process one more time.

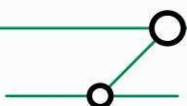
3.1 Test Cases

Identify

- Are all methods to match data properly working?
- Are all file types handled properly?
- Are all file extensions handled properly?
- Are unknown data structures handled properly?
- Is encrypted data handled properly?

Monitor

- Are all removable devices (USB, floppy disc, CD) monitored properly?
- Are all file systems monitored properly, including all special functionalities?
- Are all network protocols (Layer 2/3/4) handled properly?
- Are all intercepting network devices monitored properly?
- Is there a possibility to decrypt les using an enterprise encryption system?



- ❑ React
 - Is sensitive data blocked?
 - Are all incidents reported properly?
 - Are there reaction or blocking rules? Allow reaction rules race conditions?
 - Is there a firewall/proxy integration to block network connections?
- ❑ System Security
 - Is all sensitive traffic encrypted?
 - Exist any publicly known vulnerabilities?
 - Can vulnerabilities easily be found using simple vulnerability assessment methods?
 - Are all access rights set properly?
 - Is there a security certification like a Common Criteria Level?

4 EVALUATION

In 2008, lots of DLP solutions were released to the market. These new products, as well as new directions in DLP, are summarized by Gartner² every year. The report provides an overview on the key features of current DLP solutions. It defines also a metric on minimal requirements that have to be fulfilled by a product to be called a DLP suite. Appropriate products must provide, e.g., complete monitoring of all network data and either monitoring of data at rest or in use. The central management interface is necessary to provide a possibility to control all functions of the solution in an effective way. These requirements distinguish complete solutions from products which focus only on one part of the complete DLP approach, like content discovery.

In 2008, the report listed 16 products which met these requirements. This is an increase of almost 100% regarding the 9 solutions mentioned by Gartner in 2007. Despite this heavy growth, the market is still adolescent and evolving. This more competitive market results in more extensive capabilities of the solutions in 2008: In 2007, it was adequate to provide filtering or discovering on the network layer. In 2008, according to the evolving market, all products supported scanning of data in at least two of the three states. The integration of endpoint agents into the overall solution thereby is the most important change in product resources.

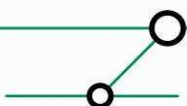
The performed evaluation focused on two products which should be examined for a customer who wants to implement a DLP solution. The key requirement was the protection of endpoint systems and their removable storage media. Employees should be able to use, e.g., USB sticks but should not be able to copy sensitive data to them. To achieve this goal, two solutions were examined in depth. The first solution is the *McAfee Host Data Loss Prevention* suite, the second one the *Websense Data Security Suite*. The Websense Data Security Suite is one of the market leading products. In contrast, the McAfee DLP suite is a quite new product.

4.1 Concrete Test Cases

Table 1: Concrete Testcases for examined DLP suites lists the test cases that were performed to evaluate whether the DLP suite were able to avoid data breaches. The performed analysis is only a subset of all tests since the focus is clearly on USB media. If the DLP suite passed a test, the corresponding cell is green, otherwise it is red. The basic text recognition is a basic check whether a simple text matching policy is applied correctly. Based on this initial check of functionality, the further tests could be performed.

So it was tested whether meta information, like the *filename* or *EXIF comments*, were also investigated. Different file types like PDF documents and word files with embedded Excel content were prepared and copied to an USB stick. These files were also compressed to add another layer of file handling. Additionally, special file system functions, like the NTFS alternate

² E. Quellet and P. Proctor. *Magic quadrant for content monitoring and filtering and data loss prevention. Technical report, Gartner RAS Core Research, 2008.*



data streams, and third party file systems, like an installed driver to access Linux partitions, were tested to be supervised properly. Another check was the copying of files to an USB hard drive: In one case, it was a really huge file of about 5 gigabyte, in the other case there were multiple partitions on the hard drive. Furthermore, it was tested whether the reaction rules are sufficient: Is data blocked or removed? If the latter, is it removed in a secure way?

Testcase	McAfee	Websense
Basic Text Recognition	Pass	Pass
Filename	Fail	Pass
PDF	Pass	Pass
Word/Excel embedment	Pass	Pass
Compression	Pass	Pass
Unknown MIME type	Fail	Fail
Metadata	Fail	Fail
NTFS Alternate Data Streams	Pass	Pass
Third party filesystems	Pass	Pass
Multiple partitions	Fail	Pass
Secure Reaction	Fail	Pass
Encryption	Fail	Fail
Fuzzing	Pass	Pass
Huge files	Fail	Fail

Table 1: Concrete Testcases for examined DLP suites

4.2 Websense Data Security Suite

To test whether the basic functionality is given, an example policy was deployed. This policy blocks files which contain the string *SECRET* from being copied to any removable storage media. Since this content policy is working, further tests could be performed. The following subsections list the most important findings of the Websense DLP suite.

4.2.1 MIME Types Not Examined In Depth

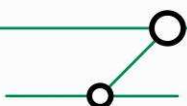
Every DLP solution supports lots of file types that can be *understood* and parsed. So a PDF document containing the test string *SECRET* was created and copied to an USB stick. Since this file was blocked correctly, the first line of the PDF document containing the words *%PDF 1.4*, which define the so called MIME type of the file, was removed to check whether a real deep content analysis is performed. This slightly modified file was not recognized by the DLP solution.

4.2.2 Missing Analysis of Meta Data

A similar test was the copying of a PNG file which contained an EXIF comment (where EXIF is a standard for embedded Meta data in image files). This EXIF comment again was the string *SECRET* and again, was not recognized by the solution.

4.2.3 Freeze During Analysis

To test the stability of the solution, a 5GB file was copied to an USB stick. During the analysis of this file, the complete system froze. After deactivating the DLP agent, the file could be copied without any problems. These freezes were reproducible and could hint to any kind of programming flaw.



4.2.4 Insufficient Encryption

The communication of the endpoint agent with the central management server is per default transported over HTTPS and thus encrypted using the SSL protocol. But since there is no sort of authentication in place, it was possible to intercept the traffic and decrypt it using a *SSL Man in the Middle Attack*. This was possible since the client does not check whether the server certificate really belongs to the correct central DLP server. So any traffic between the client and the server could be decrypted when it is possible to intercept the traffic (Figure 1: SSL MitM attack against the Websense Data Security Suite). This means that reported incidents including the blocked file can be sniffed by any attacker which is attached to the local network.

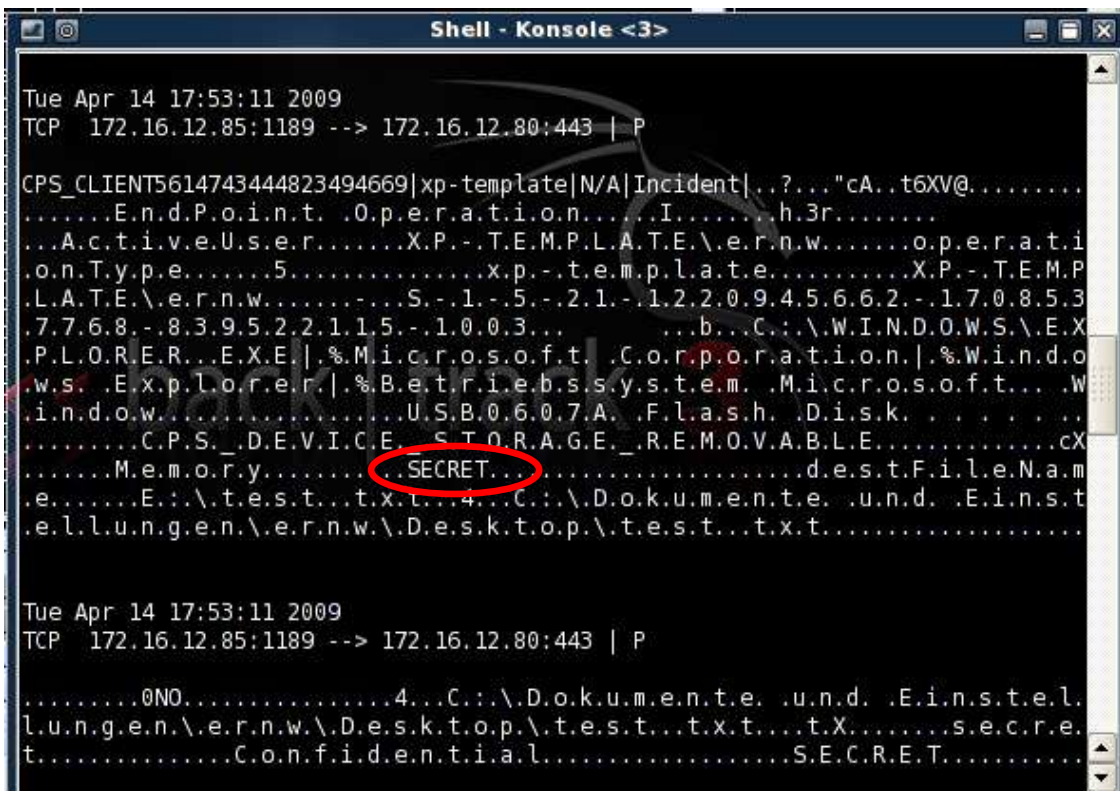


Figure 1: SSL MitM attack against the Websense Data Security Suite

4.3 McAfee Host Data Loss Prevention

Like in the previous chapter, an initial test was performed whether the basic functionality is given. Again, the text file containing *SECRET* was blocked by the example policy.

4.3.1 MIME Types Not Examined In Depth

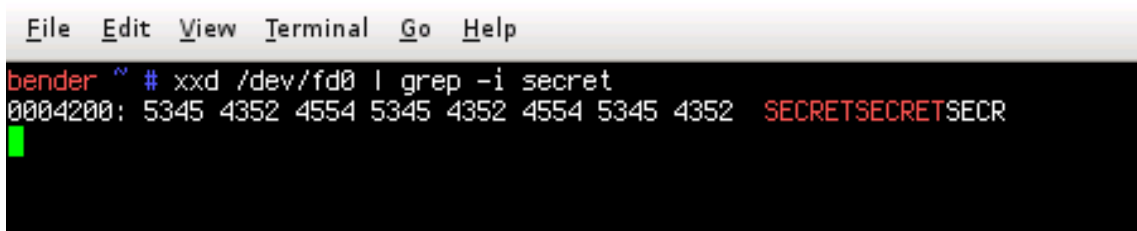
Like the Websense solution, this DLP solution also misses PDF documents without the first line – the MIME information – which identifies it to be a PDF document. This points to missing deep content analysis and thus is an insufficient MIME type recognition.

4.3.2 Missing Analysis of Meta Data

The McAfee DLP solution also misses EXIF comments in image files. But even more obvious, this solution misses also files whose *file name* is *SECRET*.

4.3.3 Insufficient React Behavior

After a file containing *SECRET* was detected, the file gets removed from the removable storage. A deep analysis of a floppy disc after this reaction showed that the DLP solution only deleted the file using standard file system methods (Figure 2: Searching for the sensitive data on a very low level). This means that the actual data is still recoverable from the removable medium and thus the reaction policy is mostly worthless.



```
File Edit View Terminal Go Help
bender ~ # xxd /dev/fd0 | grep -i secret
0004200: 5345 4352 4554 5345 4352 4554 5345 4352  SECRETSECRETSECR
```

Figure 2: Searching for the sensitive data on a very low level

4.3.4 Insufficient Monitoring of USB Hard Drives

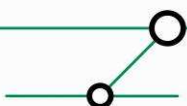
When an USB hard drive is attached, only one of the partitions of this hard drive is monitored. So it was possible to copy sensitive data to all other partitions without any further control.

4.4 Summary

The listed findings show that the examined DLP solutions are not yet matured. Even the avoidance of accidental leakage – which is the main focus of most DLP solutions – can not be assured since, e.g., not every partition of an USB hard drive is monitored. It is also possible, that DLP suites add new leakage vectors to a network: The insufficient encryption of network traffic by the McAfee DLP suite may leak all monitored incidents to attackers which are able to sniff network traffic. Such a threat also destroys the advantages that a good content scan engine may afford.

5 GENERAL PROBLEMS AND CONCLUSION

The different discovered vulnerabilities show that also software which should improve the security level of a system or a network needs a lot of development, testing and administration to be regarded as secure. The implementation of a secure development and testing process can



only be satisfied by the developer itself and should be controlled by examinations that are focused on security issues. But also the operation of such an extensive software suite needs lot of manpower and processes. So security updates must be applied, dictionaries and sensitive contents have to be updated on a regular base, and, last but not least, yet another logfile or report mail has to be read by the administrators. Additionally, both a data classification scheme and a policy how to handle sensitive data is a crucial requirement for the successful implementation of a DLP suite. If these requirements are not fulfilled, it is questionable whether a solution that increases the level of complexity in a network massively should be deployed. To roll out a DLP solution, every device which handles data – this may include almost any device – needs to be covered by the DLP suite. Otherwise, there would be always unprotected leakage vectors.

Regarding these factors, it must be carefully evaluated whether a DLP solution is an option to increase the overall level of security in a network. If all requirements are fulfilled, again, the DLP suite itself must be evaluated and examined, both in terms of operations and security.

Following ERNW's responsible disclosure guidelines, the respective vendors were contacted and the findings were reported in an appropriate way.

All names and products are property of their respective companies.

