INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Recommendation of a security architecture for data loss prevention

Luis Manuel Cerqueira Dias Pereira Ramos

Master in Information Systems Management

Supervisor:
Prof. Dr. Carlos José Corredoura Serrão, Assistant Professor
ISCTE - Instituto Universitário de Lisboa

October, 2020

TECNOLOGIAS
E ARQUITETURA

Recommendation of a security architecture for data loss prevention

Luis Manuel Cerqueira Dias Pereira Ramos

Master in Information Systems Management

Supervisor:
Prof. Dr. Carlos José Corredoura Serrão, Assistant Professor
ISCTE - Instituto Universitário de Lisboa

October, 2020

"*Imagination is more important than knowledge. For knowledge is limited, whereas imagination embraces the entire world, stimulating progress, giving birth to evolution.*"

Albert Einstein

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Carlos Serrão for his guidance, commitment and help throughout this journey. This thesis would not be possible without his valuable inputs.

To my family and friends that gave me the support and continuous motivation that I needed to achieve my goals.

Resumo

A informação e as pessoas são os ativos mais importantes de qualquer organização. A quantidade de informação que é gerada aumenta exponencialmente devido à quantidade de novos dispositivos que produzem informação. Por outro lado, cada vez mais organizações são abrangidas por algum tipo de regulamento, como o Regulamento Geral de Proteção de Dados.

As organizações implementam vários controlos de segurança, no entanto, não se focam na proteção da informação em si e a fuga da informação é uma realidade e uma preocupação crescente. Com base neste problema, existe a necessidade de proteger a informação confidencial, como dados clínicos, informação pessoal, entre outros. Neste sentido, as soluções de prevenção da fuga de informação (DLP – *Data Loss Prevention*) que têm a capacidade de identificar, monitorizar e atuar em dados considerados confidenciais, seja ao nível do *endpoint*, repositório de dados ou na rede, devem fazer parte da estratégia da segurança da informação das organizações por forma a mitigar estes riscos.

Esta dissertação vai analisar a temática da prevenção da fuga de informação e avaliar várias soluções existentes com o propósito de identificar as componentes chave deste tipo de soluções. A principal contribuição deste trabalho será a recomendação de uma arquitetura de segurança que mitigue o risco da fuga da informação e que poderá ser facilmente adaptável a qualquer solução de DLP a ser implementada pelas organizações. Por forma a comprovar a eficiência da arquitetura, a mesma foi implementada e testada para mitigar o risco de fuga da informação em cenários específicos que foram definidos.

Palavras-chave: Segurança Informática, Fuga de Informação, Confidencialidade, Dados Sensíveis

Abstract

Data and people are the most important assets of any organization. The amount of information that is generated increases exponentially due to the number of new devices that create information. On the other hand, more and more organizations are covered by some type of regulation, such as the General Data Protection Regulation.

Organizations implement several security controls, however, they do not focus on protecting the information itself and information leakage is a reality and a growing concern. Based on this problem, there is a need to protect confidential information, such as clinical data, personal information, among others. In this regard, data loss prevention solutions (DLP – *Data Loss Prevention*) that have the ability to identify, monitor and act on data considered confidential, whether at the endpoint, data repositories or in the network, should be part of the information security strategy of organizations in order to mitigate these risks.

This dissertation will study the topic of data loss prevention and evaluate several existing solutions in order to identify the key components of this type of solutions. The contribution of this work will be the recommendation of a security architecture that mitigates the risk of information leakage and that can be easily adaptable to any DLP solution to be implemented by organizations. In order to prove the efficiency of the architecture, it was implemented and tested to mitigate the risk of information leakage in specific proposed scenarios.

Keywords: Information Security, Data Loss Prevention, Confidentiality, Sensitive Data

Table of Contents

Acknowledgements ..... **i**

Resumo ..... **ii**

Abstract ..... **iii**

List of Tables ..... **vii**

List of Figures ..... **viii**

Abbreviations and Acronyms ..... **xi**

Chapter 1 – Introduction **1**

1.1. Background 1

1.2. Research Problem and Motivation 2

1.3. Objectives and Research Questions 3

1.4. Research Methodology 3

1.4.1 Problem identification and motivation 4

1.4.2 Define the objectives for a solution 4

1.4.3 Design and development 4

1.4.4 Demonstration 5

1.4.5 Evaluation 5

1.4.6 Communication 5

1.5. Document Structure 5

Chapter 2 – Literature Review **7**

2.1. Introduction to Data Loss Prevention 7

2.2. Traditional Security Technologies 8

2.2.1. Intrusion Detection/Intrusion Prevention System 8

2.2.2. Firewalls 9

2.2.3. Antimalware 10

2.2.4. Virtual Private Networks 11

2.3. Data Breaches 12
2.3.1. The motivations behind attacks 13
2.3.2. Costs of a data breach 14
2.4. Data Loss Vectors 16
2.4.1. Data at Rest 17
2.4.2. Data in Motion 17
2.4.3. Data in Use 18
2.5. Data Classification 18
2.6. Detection Technologies 21
2.7. Data Loss Prevention Solutions 23
2.7.1. Digital Guardian 25
2.7.2. Forcepoint 26
2.7.3. Intel Security 26
2.7.4. OpenDLP 27
2.7.5. Symantec 27
2.7.6. DLP Solutions Comparison 28
2.8. Conclusion 30
Chapter 3 – Proposed Solution and Implementation 31
3.1. Introduction 31
3.2. Proposed Solution High Level Design 32
3.2.1 High Level Design Components 33
3.3. DLP Implementation 36
3.3.1 Proposed Architecture Mapped to Symantec Data Loss Prevention 37
3.3.2 Installation Tiers 39
3.3.3 System Requirements for Test Environment 40
3.3.4 Solution Pack 41

3.3.5 DLP Agent Configuration and Installation 42
3.3.6 Policies 49
3.4. Conclusion 52
Chapter 4 – Testing and Validation 53
4.1. Introduction 53
4.2. Test Cases 53
4.2.1. Test Case 1 – PCI DSS Data 53
4.2.2. Test Case 2 – Data Classification Policy 57
4.2.3. Test Case 3 – Personal Data 60
4.2.4. Test Case 4 – Custom Detections 64
4.3. Conclusion 71
Chapter 5 – Conclusion 72
5.1. Conclusion 72
5.2. Future Work 73
Appendix A: Installing Symantec Data Loss Prevention 75
References 87

List of Tables

Table 1 - Data Classification Schemes 19

Table 2 - DLP Product Comparison 29

Table 3 - Solution Packs 41

Table 4 - DLP Agent Installation Files 45

Table 5 - DLP Agent Command-Line Arguments 46

List of Figures

Figure 1 - Design Science Research Methodology (DSRM) Process Mode 4

Figure 2- Distribution of the benchmark sample by root cause of the data breach ... 13

Figure 3 – Per capita costs of a data breach..... 14

Figure 4 - Per capita cost for three root causes of the data breach 15

Figure 5 - Relationships between mean time to identify and average total cost 15

Figure 6 - Days to identify and contain a data breach by industry sector..... 16

Figure 7 - A taxonomy of DLP solution 17

Figure 8 - 2017 Gartner Magic Quadrant for Enterprise Data Loss Prevention..... 24

Figure 9 - DLP High Level Design..... 32

Figure 10 - DLP for Email Single MTA Architecture..... 34

Figure 11 - DLP for Email using multiple MTA Architecture 34

Figure 12 - DLP for Web Architecture 35

Figure 13 - DLP for Network Architecture 36

Figure 14 – Proposed Architecture Mapped to Symantec Data Loss Prevention..... 38

Figure 15 – Single tier DLP deployment 39

Figure 16 – Two-tier DLP deployment..... 39

Figure 17 – Three-tier DLP Deployment..... 40

Figure 18 - DLP Agent Configuration 44

Figure 19 - DLP Agent Packaging..... 45

Figure 20 - DLP Agent in Windows Task Manager 48

Figure 21 - Policy PCI DSS 54

Figure 22 – PCI DSS Discovery Scan 54

Figure 23 - PCI DSS Discovery Scan Conetnt 55

Figure 24 - PCI DSS Discovery Scan Result..... 55

Figure 25 – PCI DSS Discovery Scan Result Details..... 56

Figure 26 - PCI DSS Discovery Incident Detail 56
Figure 27 - Data Classification Policy 58
Figure 28 - Data Classification Policy Response 58
Figure 29 - Data Classification Policy Incident Detail 58
Figure 30 - Data Classification Policy Endpoint Block 59
Figure 31 - Data Classification Policy Discovery Scan 59
Figure 32 - Personal Data Index 61
Figure 33 - Personal Data Indexed Files 61
Figure 34 - Personal Data Policy 61
Figure 35 - Personal Data Policy Response 62
Figure 36 - Personal Data Policy Web HTTP Block 62
Figure 37 - Personal Data Policy Web HTTP Monitor 63
Figure 38 - Personal Data Incidents 63
Figure 39 - Personal Data Discovery Scan 64
Figure 40 - Portuguese ID Data Identifier 65
Figure 41 - Portuguese National ID Card Policy 67
Figure 42 - Portuguese National ID Card Policy, Response Rule 67
Figure 43 - Portuguese National ID Card Incident 67
Figure 44 - Portuguese Social Security Number Data Identifier 68
Figure 45 - Portuguese Social Security Number Policy 70
Figure 46 - Portuguese Social Security Number Incident 70
Figure 47 - DLP Install, Step 1 75
Figure 48 - DLP Install, Step 2 75
Figure 49 - DLP Install, Step 3 76
Figure 50 - DLP Install, Step 4 76
Figure 51 - DLP Install, Step 5 77

Figure 52 - DLP Install, Step 6 77
Figure 53 - DLP Install, Step 7 78
Figure 54 - DLP Install, Step 8 78
Figure 55 - DLP Install, Step 9 79
Figure 56 - DLP Install, Step 10 79
Figure 57 - DLP Install, Step 11 80
Figure 58 - DLP Install, Step 12 80
Figure 59 - DLP Install, Step 13 81
Figure 60 - DLP Install, Step 14 81
Figure 61 - DLP Install, Step 15 82
Figure 62 - DLP Install, Step 16 82
Figure 63 - DLP Install, Step 17 83
Figure 64 - DLP Install, Step 18 83
Figure 65 - DLP Install, Step 19 84
Figure 66 - DLP Install, Step 20 84
Figure 67 - DLP Install, Step 21 84
Figure 68 - DLP Install, Step 22 85
Figure 69 - DLP Install, Step 23 85
Figure 70 - DLP Install, Step 24 86
Figure 71 - DLP Install, Step 25 86

Abbreviations and Acronyms

DLP – Data Loss Prevention

FTP – File Transfer Protocol

HTTP – Hypertext Transfer Protocol

HTTPS – Hypertext Transfer Protocol Secure

ICAP – Internet Content Adaptation Protocol

IPSec – Internet Protocol Security

GDPR - General Data Protection Regulation

MTA – Message Transfer Agent

NAS - Network-attached storage

OCR - Optical character recognition

OSI - Open Systems Interconnection

PCI DSS – Payment Card Industry Data Security Standard

SOX – Sarbanes–Oxley

SSL – Secure Sockets Layer

Chapter 1 – Introduction

1.1. Background

In the past, all organizations had what was considered as the perimeter of the network, in which there was a clear separation of the organization's secure network and the outside world. The information was, to some extent protected, since it was stored within the secure perimeter of the organization and the only way to access that information was to be physically in the network (O'Hanley & Tiller, 2013).

Nowadays, in the digital economy, information flows at a high speed and mobility and cloud trends translate into greater productivity, since it is possible to access information from anywhere (Rocha et al., 2015). However, the secure perimeter of the organization becomes broader, which raises information security risks, since it is extremely difficult to control information flows when most security controls are implemented within the organization's network. At the same time, the amount of information is increasing exponentially and according to Reinsel, Gantz and Rydning (2018) it will grow from 45 zettabytes in 2019 to 175 zettabytes in 2025, whether it is created, captured or replicated.

Organizations keep confidential information of customers, partners and information that may be related to some type of regulation, such as General Data Protection Regulation (GDPR). In addition, many companies are victims of information leakage, which translates into an impact on one's reputation, competitiveness, and often on financially heavy fines (Shabtai, Elovici & Rokach, 2012). Data leakage occurs when confidential data falls into unauthorized hands. This data includes intellectual property, financial data, patient data, personal credit card data and other confidential information, depending on the business and the industry. This is an important issue for companies, as the number of incidents and the cost to those who face them continue to increase. Whether caused by malicious intent or an inadvertent error, whether internal or external, the exposure of confidential information can seriously harm an organization (Baby & Krishnan, 2017).

To safeguard information and to mitigate the risk of data loss, Data Loss Prevention (DLP) solutions can be implemented. DLP is a solution for detecting and preventing information leaks from within an organization’s network (Alzhrani et al., 2016), as such, it can identify different types of data that are valuable to organizations, monitor different channels for data leakage (data loss vectors), including endpoint, email, web, data

repositories and act upon incidents that can occur by applying an action such as blocking an email, alert the user or notify a security analyst.

1.2.Research Problem and Motivation

Several studies have been published around data breaches that shows this is an increasing problem within the information security realm. A study from Verizon (2018) reiterate that the majority of data breaches are perpetrated by outsiders followed by internal actors. The most affected companies were healthcare organizations and small businesses. On the other hand, a study by Ponemon Institute (2017) shows that companies had larger data breaches during 2017. Data is one of the most important organizational intangible assets and therefore its protection should be a priority.

There are different security controls that organizations implement to mitigate information security risks. These security controls are put in place to adopt the primary principles of information security: confidentiality, integrity and availability (Andress and Winterfeld, 2014). Although a number of security controls are implemented, such as, firewalls, intrusion detection systems, antimalware, data leakage still occurs (Tahboub and Saleh, 2014).

In addition, Alneyadi, Sithirasenan and Muthukkumarasamy (2015) agree that data loss prevention solutions are increasingly being implemented by organizations to protect against data loss since they are able to use the content of files to detect and prevent unauthorized access to sensitive information.

Moreover, every DLP solution that exist have specific modules that make the overall solution to mitigate the risk of data loss in different vectors, however, each solution uses a different naming convention for the different modules and it is not clear which components are key to protect information in different states. This study aims to standardize the key components and build a generic security architecture.

The motivations for this master thesis, is that the information is "the new oil". Digital transformation gives organizations agility and simplification of business processes, however, they increase risk, loss and / or leakage of data, since access to it has become trivialized. Information is then one of the main assets of each company and several studies have been published around the increase in data loss incidents. These factors contribute

for the relevance of this theme and the purpose of this dissertation is to mitigate these risks with the implementation of solutions to prevent information leakage.

Furthermore, my professional background allowed me to be part of the implementation of data loss prevention projects in large corporations across Europe, Middle East and Africa (EMEA), therefore, is a subject of interest.

1.3.Objectives and Research Questions

The goal of this thesis is to propose a security architecture that will reduce the risk of data loss and to mitigate the problems described in the previous section. On the other hand, the proposed architecture will focus in the key components that address data loss so it can be easily adapted to any existing data loss prevention solution.

In order to achieve this, a scientific literature review around the subject will be conducted that will help identify how these solutions work and what are the main components that needs to be deployed. Multiple data loss prevention solutions, commercial and open source will both be considered and studied to understand the different options available.

The main research question which this work attempts to answer is “Can a generic data loss prevention security architecture mitigate the risk of data loss?”

To answer this question, the following objectives are proposed:

1. Study, definition and evaluation of a data loss prevention architecture that is generic and adaptable to different DLP solutions.
2. Identification of the key components in the data loss prevention architecture that mitigate the risk of data loss
3. Development of DLP Policies to identify personal data and confidential information.

1.4.Research Methodology

The Design Science Research Methodology (DSRM) is the most appropriate methodology to address the research question and to achieve the goals of this thesis. Hevner et al. (2004) states that design science research (DSR) “creates and evaluates IT

artifacts intended to solve identified organizational problems". IT artifacts are made up of constructs, models, methods, and instantiations (Hevner et al., 2004).

Hevner et al. presented a process and guidelines for design science research within the discipline of Information Systems.

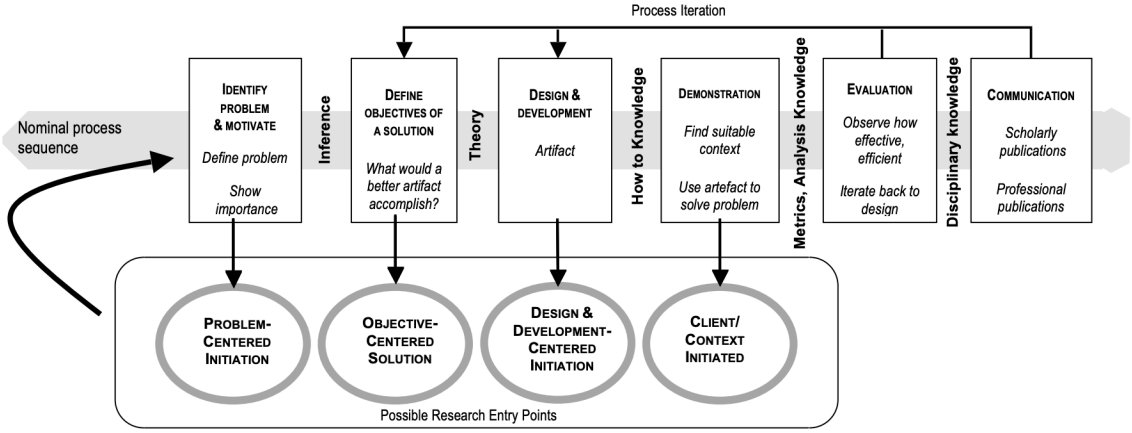

Figure 1 - Design Science Research Methodology (DSRM) Process Mode

Figure 1 source (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007).

1.4.1 Problem identification and motivation

This is the first step as per the DSRM. The problem identification is related to the research question "Can a generic data loss prevention security architecture mitigate the risk of data loss?". Both problem identification and motivation are described in section 1.2.

1.4.2 Define the objectives for a solution

The objectives are described, and recommendations will be made as a result of the proposed objectives in section 1.3. As a summary, the objectives are:

- Study, definition and evaluation of a data loss prevention architecture that is generic and adaptable to different DLP solutions.
- Identification of the key components and features of a DLP solution based on the recommended architecture.
- Development DLP policies to identify sensitive data.

1.4.3 Design and development

The artifact will be designed and developed based on the defined objectives. This results from the interpretation and understanding of the scientific research and from

existing data loss prevention solutions documentation. Use cases will be presented with the goal of being addressed and solved based on the research objectives.

1.4.4 Demonstration

The artifact will be implemented to confirm that it will be effective protecting organizations from data loss. A commercial DLP solution will be used to configure the use cases identified in the design and development phase.

1.4.5 Evaluation

In the evaluation phase we assess whether the artifact supports the proposed solutions to the problems. The proof of the artifact will be analyzed and observed, evidences will be collected in a proof-of-concept to support the solution and as a validation of the method.

1.4.6 Communication

The outcome of the thesis would be presented, and the results shared through the master thesis document.

1.5. Document Structure

The current thesis is composed by five chapters which are intended to reflect the different stages up to their completion.

The first chapter introduces the research background, problem identification and motivation, research question and objectives, methodology used and a description of the document structure.

The second chapter reflects the framework, described as literature review that contains the reading review for this chapter as well as the challenges that justify the implementation of Data Loss Prevention solutions. It describes the different states of data, detection technologies used, and it will also analyze multiple data loss prevention solutions.

The third chapter is focused on the design and development of our artifact. To achieve this, a DLP solution will be implemented based on the proposed architecture, including the necessary endpoint agent to solve the proposed objectives.

The fourth chapter will demonstrate the efficacy of the artifact and the configuration of the use cases as a validation of the proposed solution.

The fifth and final chapter presents the conclusions of this study, recommendations and future work.

Chapter 2 – Literature Review

The goal of this chapter is to introduce the reader to the concepts of Data Loss Prevention: what is it, how it works, why is needed and to explain how DLP solutions works with its different modules.

2.1. Introduction to Data Loss Prevention

Cisco (2019) describes data loss prevention or DLP as a set of technologies, products and techniques that are designed and built to protect sensitive information and to prevent the information from leaving the organization. Data leakage is defined as an accidental or unintentional distribution of private or sensitive data to an unauthorized entity (Shabtai, Elovici and Rokach, 2012). DLP solutions work by identifying what confidential information is (either through the use of regular expressions, indexes, among others) in different vectors, such as, endpoint, network, email, web or cloud with the goal to stop information, such as personal records, intellectual property, financial data, from being sent, either accidentally or intentionally outside the corporate network (CISCO, 2019).

Organizations use many security controls; however, they are not data centric whereas DLP solutions focus on the data. Some of the benefits of deployment a DLP solution are (Noble et al., 2010):

- **Protecting business data and intellectual property**: One of the main benefits of DLP is to protect the information that is important for the company. Organizations handles a large amount of information, such as customer information, health information, source code and DLP will support in keeping this information safe.
- **Compliance**: DLP help an organization to meet regulatory requirements protecting data that falls into a specific regulation. Most DLP solutions have templates pre-configured to support addressing these needs.
- **Reduce data breaches**: By reducing the risk of data loss, the financial risk for an organization, decreases.
- **Training and awareness**: Most organizations have written security policies but still, it is difficult to collect evidences that they are in fact being followed. DLP solutions can alert an end-user by the use of popups (endpoint) or emails each time a policy is violated. This is key to keep the users educated about data security policies and what can be done with the information.

2.2.Traditional Security Technologies

There are a number of security controls that are typically implemented to increase the overall security posture; however, they lack some of the key capabilities of data loss prevention solution that analyze the contents of the data. According to Tahboub and Saleh (2014), the traditional approach to security, such as firewalls, do not protect against data loss, moreover, intrusion detection systems (IDS), intrusion prevention systems (IPS) and antimalware technologies are solutions that work in conjunction with data loss prevention systems. One of the most important assets of organizations is data. Therefore, protecting this data should be the first priority. Although companies have security measures, data leakage still occurs. This leak occurs when confidential data is revealed to unauthorized parties, whether intentionally or not. The loss of confidential or sensitive data can seriously affect a company's reputation, the trust of customers and employees, competitive advantage and, in some cases, lead to the closure of the company.

The traditional security technologies implemented are described below.

2.2.1. Intrusion Detection/Intrusion Prevention System

Intrusion detection system (IDS) and intrusion prevention system (IPS), is a software that monitors the network or systems for malicious activities. It can be deployed in a variety of ways, such as virtual machine or dedicated hardware. The difference between an IDS and an IPS is that an IDS it is not in-line with the network traffic whereas an IPS is in-line with network traffic which can not only detect but prevent malicious activities (Ashoor and Gore, 2011).

According to Scarfone and Mell (2007), intrusion prevention systems can be classified in four types:

- Network-Based: Monitors network traffic and analyzes the network and application protocol activity in order to identify malicious activity.
- Wireless: Focusing on wireless protocol, this type of IPS monitors the traffic to identify suspicious activity.

- Network Behavior Analysis (NBA): Examines network traffic to identify threats that generate unusual traffic flows such as port-scanning, distributed denial of service attacks, among others.
- Host Based: Monitors the activity in a single host, by installing an agent that identifies suspicious activity.

Liao et al. (2013) identified that intrusion Detection and Prevention System uses different detection methodologies that can be classified as three major categories:

- Signature-based: Monitors network traffic and compares it with known threat signatures.
- Anomaly-based: Monitors network traffic and compares it with a baseline to identify potential deviations. It is common to monitor regular activities such as network connections, hosts or users during a specific time period to develop the baseline.
- Stateful Protocol Analysis: This category of IPS identifies deviation of protocol states by comparing with protocol standards from international standard organizations.

2.2.2. Firewalls

Cisco (2017) describes a firewall as being a network device that monitors incoming and outgoing network traffic and decides whether to allow or to block specific traffic based on a defined set of security rules. There are different types of firewalls that can be categorized in four different types (Moraes, 2011):

- Packet Filters: Packet filtering firewalls work on the basis of rules defined by access control lists. All the packets are checked against the rules defined to determine what action should be applied. This type of firewall is stateless because they don't have the concept of state table.
- Circuit-Level Proxies: This type of firewall is deployed at the session layer of the OSI model and they monitor TCP three-way handshake to see if a requested connection is legitimate or not.
- Application-Level Proxies: Also called as proxy firewalls, they inspect data packets at the application level to filter incoming traffic between the local network and the source of the traffic. Instead of letting traffic connect directly,

the proxy firewall first establishes a connection to the source of the traffic and inspects the incoming data packet. Application-level proxies can also be configured as caching servers which increases the network performance.

- Stateful Firewalls: This type of firewall adds the concept of connections and states for packet filter implementations. They come with both packet inspection technology and TCP handshake verification to create a greater level of protection. For access control rules, they use a group of packets belonging to the same connection rather than individual packets.

2.2.3. Antimalware

Antimalware software also known as antivirus is a software that detects and removes malicious software and, according to Koret and Bachaalany (2015), gives better protection than the one offered by the underlying operating system.

Malware can be broadly classified into different categories (Vinod et al., 2009; Oriyano, 2016)

- Viruses: Viruses are the best know form of malicious software; when executed it replicates itself and can modify another executable file.
- Worms: Worms are self-replicating programs. They replicate in order to spread to other computers.
- Spyware: Spyware is a type of malicious software that gathers information about the user, such as webpages frequently visited, keystrokes, etc. They are usually stealthy and installed when free or trial software is downloaded.
- Adware: Adware works by displaying advertisements after the malicious software is installed. It can replace home pages in browsers and display pop-ups. Adware usually comes embedded with free software.
- Trojans: Trojans or trojan horse are a special type of malware that emulate the behavior of a real program to damage to the computer. They usually spread with social engineering techniques that manipulates people in order to install the malicious software.
- Botnet: Botnets consists on a network of connected computers that have been infected by a worm or a trojan. Botnets are usually used to send spam and perform denial of services attacks.

Furthermore, there are three main malware detection techniques (Vinod et al., 2009; Idika and Mathur, 2007):

- Signature-based: This technique consists in creating a signature (sequence of bytes within the code) of the malicious software than antimalware solutions use to compare with files being executed. Signature-based antimalware solutions can only detect for which a signature exists.
- Specification-based: This detection technique produces a low rate of false alarms; instead of trying to approximate the implementation of a system or application, specification-based detection attempts to approximate the requirements for an application or system. With this methodology, manually developed specifications are used to characterize legitimate program behaviors. Since this method is built with legitimate behaviors, it does not generate false alarms when unusual, however, legitimate program behaviors are encountered.
- Behavior-based: Behavior-based detection works by evaluating an object and what it is trying to do before it can execute that behavior. It usually occurs in two phases: training and detection. An advantage of this technique is that it is able to detect zero-days attacks, in other words, it can block malicious software exploiting a known vulnerability to which the software vendor didn't released a patch, however, it can generate false-positives if the object exhibits behaviors not seen in the detection phase.

2.2.4. Virtual Private Networks

A virtual private network is an encrypted connection over the internet from a network or device which is widely used in corporate environments (Cisco, 2018).

There are two types of virtual private networks: Remote Access VPN and site-to-site VPN (Jaha, Shatwan and Ashibani, 2008; Cisco, 2018). For the goal of this thesis, it is most important to describe remote access VPN which allows a user to remote access internal resources within the organization.

Remote access VPN allows a user who is working remotely, to securely access data and applications that are in the corporate datacenter. It encrypts all traffic sent and received for additional security and eliminates the need to have dedicated lines from the

internet service provider. According to Lakbabi, Orhanou and Hajji (2012), the most common technologies used in VPNs are IPSec and SSL. SSL VPN are common in clientless architectures where a remote access VPN is established through the user web browser and IPSec requires a client software to be established. IPSec is a standard commonly used to implement VPNs that enables the protection of all types of Internet protocol (IP) communications by protecting multiple peers at the network layer, in both the IPv4 and IPv6 environments (Adeyinka, 2008).

2.3.Data Breaches

Before describing the types of data loss or data breaches, it is important to know what a data breach is. A data breach is a security incident that involves the intentional or unintentional access, disclosure, manipulation or destruction of data (Fowler, 2016). In addition, Verizon (2018) describes a data breach as an incident that results in the confirmed disclosure, not just potential exposure, of data to an unauthorized party.

Several studies have been done around data breaches, which shows an increasing number of breaches occurring, Garrison and Ncube (2011) presented, information to companies and individuals about the possible correlation between types of data breaches and their companies. This study aims to increase the knowledge about data breaches in a five-year timeline. Data have been classified and analyzed by type of breach, company, and size of the records. The conclusions were that educational institutions are more likely to suffer a data breach and the cause is associated to type of hacker and exposed. The proportion of insider incidents is smaller than the other breach types and the number of records breached is independent of institution and breach type. This study also mentions that the knowledge around the data breaches characteristics and the relationship between type of breach and company will allow a more effective protection of classified information.

In addition, Layton and Watters (2014) argue that a key question for business is to determine the potential cost of a data breach as this will help assess the risk that data breaches pose, most importantly, for financial results. Although many studies in this area are dependent of subjective data (research and interviews), the authors resort to objective case studies to adjust the parameters of a general model of data breach costs, derived from applied econometrics. This helps to triangulate the findings of previous articles, but it also

overcomes some of the research limitations, especially the self-selection bias that may occur. In addition, the results allow interested parties to reproduce them, including changing any numbers that may be necessary for their own circumstances. Although many studies have identified intangible costs for companies as the main source of variation in costs, the main finding is that tangible costs are very significant in themselves. They argue that, regardless of whether tangible or intangible costs are the biggest or smallest contributors, companies need to focus on the total cost of the bottom line and implement measures to reduce the risk of data breaches in the most economical way.

Furthermore, Thomas et al. (2017) developed a study around data breach. During the period from March 2016 to March 2017, 788.000 potential victims were identified, 12.4 million potential victims of phishing and 1.9 billion usernames and passwords exposed through data breaches, negotiated on black market forums. Using this data set, they explored the extent to which stolen passwords, which originate from thousands of online services, allow an attacker to obtain a victim's valid email credentials and thus complete control of their online identity. The authors showed how to strengthen authentication mechanisms to include additional risk signals, such as geolocation of user history and device profiles, helping to reduce the risk of theft.

2.3.1. The motivations behind attacks

There are a number of reasons why a data breach can occur, according to a study made by Ponemon Institute (2018), there are three main causes: malicious or criminal attacks, system glitch or human error. A study from 2018 gathers data from 477 organizations and interviews were made to more than 2.200 individuals knowledgeable about the data breach incidents.

The following figure describes the distribution of the root causes of a data breach.

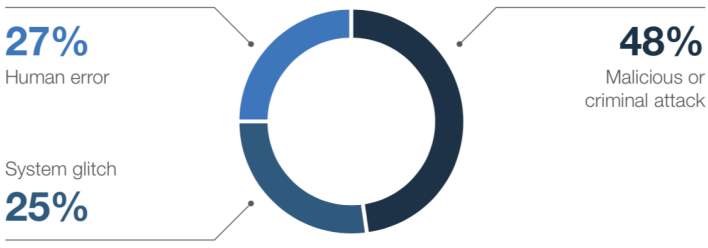

Figure 2- Distribution of the benchmark sample by root cause of the data breach

Figure 2 source: Ponemon Institute (2018)

Malicious attacks caused by hackers or criminal insiders that can be an employee, contractor or a third party, contributed for 48% of incidents while human error is due to negligent employees who caused an incident because of their carelessness or contractors represented 27%; System glitch that includes business and IT process failures represented 25%.

2.3.2. Costs of a data breach

The costs of a data breach are higher in 2018 when comparing to 2017; the global cost of a data breach increased by 6.4 percent and the per capita cost increased by 4.8 percent. The average size of a data breach, which essentially means the number of records lost or stolen also increased by 2.2 percent.

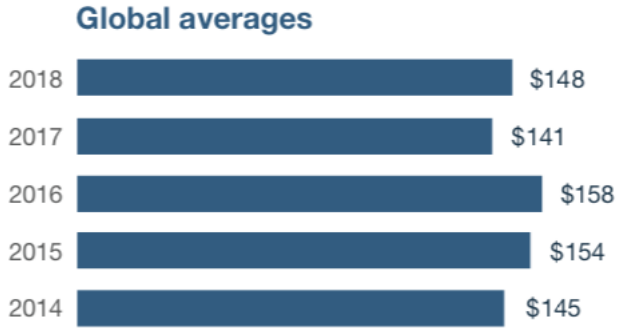

Figure 3 – Per capita costs of a data breach

Figure 3 source: Ponemon Institute (2018)

Malicious or criminal attacks have the highest costs as per Figure 4 making a total of $157; in 2018 the cost of a data breaches due to human error or system glitch was $128 and $131, respectively.

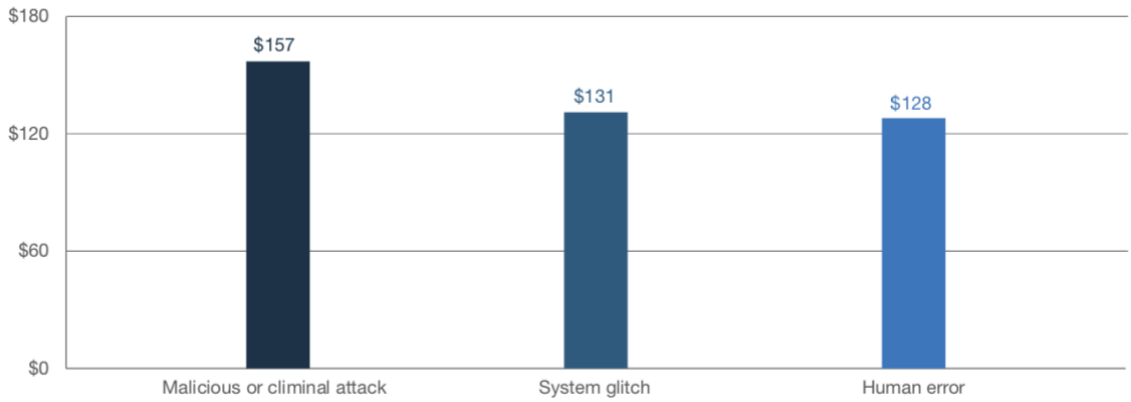

Figure 4 - Per capita cost for three root causes of the data breach

Figure 4 source: Ponemon Institute (2018)

It is important to note that, one of the key factors that influences the costs of a data breach is the mean time to identify (MTTI) and the mean time to contain (MTTC). From the sample of 477 companies, the MTTI was 197 days and the MTTC was 69 days. Companies that are able to contain a breach in less than 30 days saved over $1 million comparing to those that took more than 30 days to resolve the incident (Ponemon Institute, 2018).

IBM (2019) presented an updated study on the cost of data breaches, and mean time to identify increased to 206 days whereas the mean time to contain increased to 73 days.

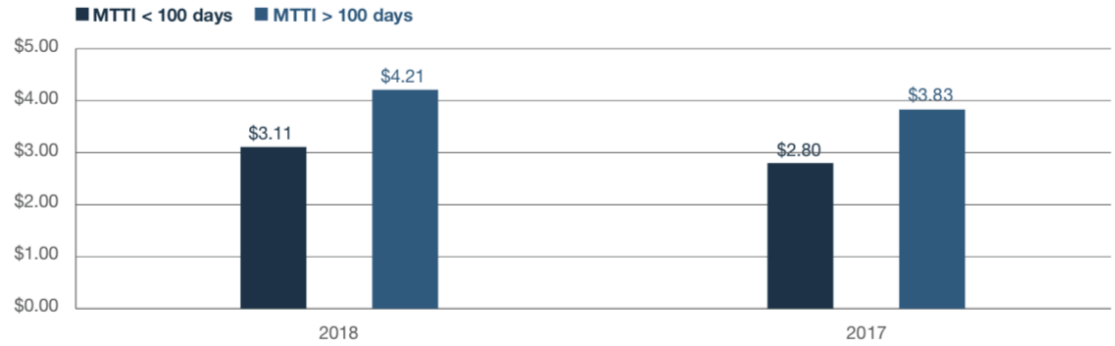

Figure 5 - Relationships between mean time to identify and average total cost

Figure 5 source: Ponemon Institute (2018)

On the other hand, companies that invest in incident response teams and use encryption solutions are also able to reduce costs. An organization can save up to $14 per compromised record when having a dedicated incident response team and $13 per capita when using extensively encryption solutions.

According to IBM (2019), healthcare and public take the most time to identify and contain and financial sector takes the least time to respond to a data breach.

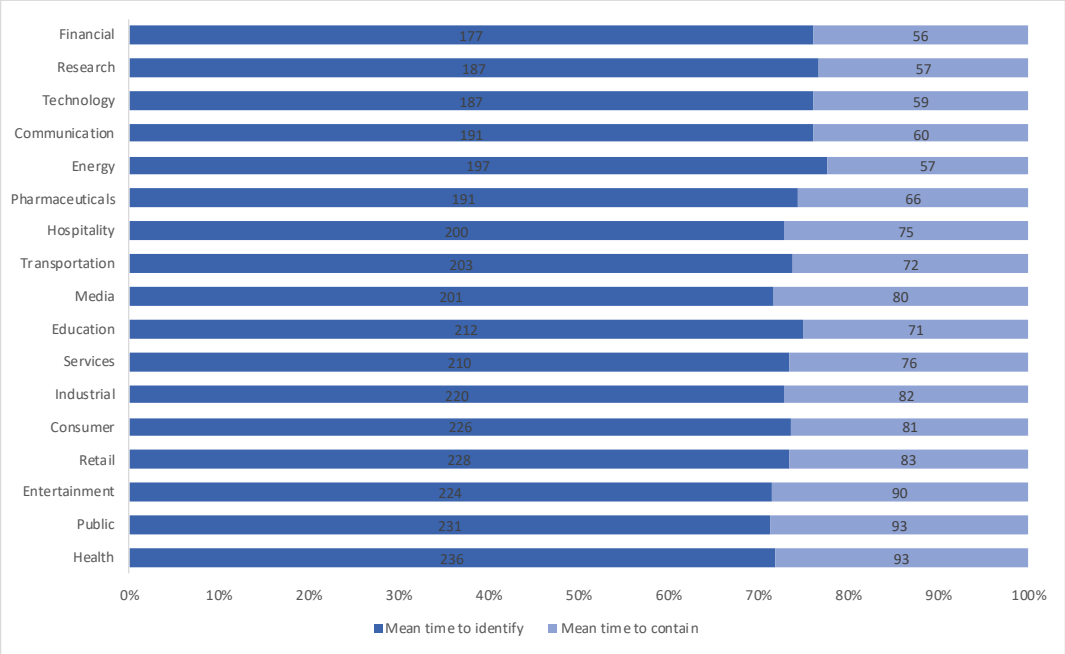

Figure 6 - Days to identify and contain a data breach by industry sector

Figure 6 source: IBM (2019).

2.4.Data Loss Vectors

According to Shabtai, Elovici and Rokach (2012), DLP solutions can be described according to a taxonomy that includes the following attributes: data state, deployment scheme, leakage handling approach and actions taken once a data leakage occurs.

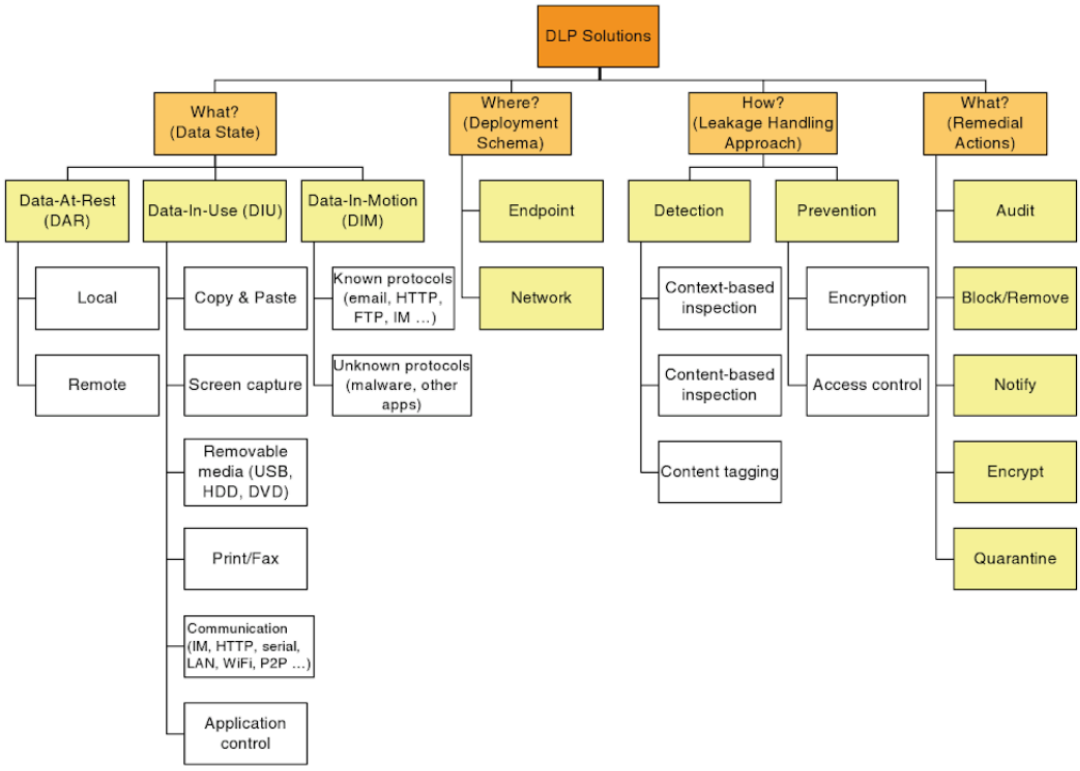

Figure 7 - A taxonomy of DLP solution

Figure 7 source: Shabtai Elovici and Rokach, 2012.

Information generally exits in the following states: data at rest, data in motion and data in use. Each state is addressed by different components within the DLP solution.

2.4.1. Data at Rest

Data that usually resides in filesystems, storage area networks (SANs) or databases. Furthermore, Wühner and Pretschner (2012) describes DLP solutions that protect data at rest identify sensitive data in persistent storage locations. Data loss prevention solutions address data at rest by using a crawler. A crawler essentially will integrate with the target being scanned such as a filesystem or website to identify confidential information. Once the information is identified, the files are opened and compared to the DLP policies. It may also allow to schedule a scan against the targets configured to automatically identify confidential information in different systems or solutions.

2.4.2. Data in Motion

In order to identify data travelling across an organization, DLP solutions uses this component to analyze network traffic. Data in motion means that information moves within the network to the outside world, whether via email, instant messaging, peer-to-

peer or other communication mechanisms (Liu and Kuhn, 2010). There are multiple ways for a DLP solution to integrate with the network. It can be either passively using a network span or they can integrate with proxies or email gateways to inspect web and email information, respectively. Depending on the use case, many integrations can be required. Depending on the policies created, DLP solution has the capability to alert, block or apply different actions in real-time (assuming an in-line configuration). If a user sends an email, the DLP can integrate with the email gateway and if a policy is violated, the email can be blocked, encrypted, or modified to be processed in the different way within the email stream. Organizations can also deploy a passive component to receive a copy of the traffic and match it against the DLP policies, this scenario as stated before, usually works by deployment a network tap or span (port mirroring) being limited to identity confidential information but without the ability to block it.

2.4.3. Data in Use

Data in this state refers to information currently used at the endpoints such as http, https, print, copy to external storage (Kaur, Gupta and Singh, 2017). In order to actively monitor data in use it is required to install an agent provided by the DLP solution that reports to the central management server. Some limitations exist on the endpoint depending on the rules and policies created; some policies might not apply to the endpoint and therefore it is important to understand the use cases that will be addressed in the endpoint.

2.5.Data Classification

Data Loss Prevention solutions work by identifying potential data leakages by monitoring, detecting and blocking confidential information either while in-use (endpoint), in-motion (network) or at-rest (storage). In order to identify sensitive information, some mechanisms are used, such as exact data matching, structured data fingerprinting, statistical methods (Bayesian and machine learning), regular expression, lexicons, keywords and watermarks (Ouellet, 2009). Data is classified as structured or unstructured. Structured data typically resides in fixed fields such as a spreadsheet or database while unstructured data refers to text documents, PDF files, video or audio (Gandomi and Haider, 2015).

Hence, it is important that organizations proper classify their information so DLP solutions can be more efficient. The primary objective of data classification is to formalize the process of securing data based on the assigned label of importance; classifying data "is the primary means by which data is protected based on its need for secrecy, sensitivity, or confidentiality" (Stewart, Chapple and Gibson, 2015, p. 18). Not all data is the same and different security controls should be in place to protect confidential information.

According to Harris (2010), "there is no hard and fast rules on the classification levels that an organization should use" (p. 110). Some organizations may use two layers of classification while another company may choose to use four. The next table explains the types of classifications available, however, note that some classifications are more commonly used in military whereas others are used for private sector (commercial business).

Table 1 - Data Classification Schemes

Classification	**Definition**	**Examples**	**Organizations that would use this**
Public	• In case of information disclosure, it would not cause an adverse.	• General available information • Upcoming projects	Commercial business
Sensitive	• Requires special precautions to ensure the integrity and confidentiality of the data by protecting it from unauthorized modification or deletion. • Requires higher than normal assurance of accuracy and completeness.	• Financial information • Details of projects • Profit earnings and forecasts information	Commercial business

Private	• Personal information for use within a company. • Unauthorized disclosure could adversely affect personnel or the company.	• Work history • Human resources Information • Medical information	Commercial business
Confidential	• For use within the company only. • Unauthorized disclosure could seriously affect a company.	• Trade secrets • Healthcare information • Source code • Competitive information about the company	Commercial business Military
Unclassified	• Data is not sensitive or classified.	• Computer manual and product brochures • Recruiting information	Military
Sensitive but unclassified (SBU)	• Minor secret. • If disclosed, it may not cause serious damage.	• Medical data • Answers to test scores	Military
Secret	• If disclosed, it could cause serious damage to national security.	• Troops placement plans • Nuclear bomb placement	Military
Top secret	• If disclosed, it could cause grave damage to national security.	• Blueprints of new wartime weapons • Spy satellite information	Military

		• Espionage data	

The following are the most common levels of sensitivity from the highest to the lowest for private sector:

- Confidential
- Private
- Sensitive
- Public

The following are the most common levels of sensitivity from the highest to the lowest for military use:

- Top Secret
- Secret
- Confidential
- Sensitive but unclassified
- Unclassified

2.6.Detection Technologies

One of the key features that Data Loss Prevention Solutions perform is content analysis. Regardless the state of the data, whether in motion, in use or at rest, DLP must be able to analyze the contents of the files and understand whether there is confidential information.

Several content analysis techniques are commonly available within DLP Solutions:

- **Regular expressions**: Regular expressions is one of the most common content analysis techniques for specific rules, such as, identifying 16-digit credit card, passport numbers, citizen numbers, among others. However, regular expressions may generate false positives, such rules need to be fine-tuned to avoid these scenarios. As an example, if an organization wants to detect 16-digit credit card numbers, the rule much validate whether the checksum is valid and not only detect a 16-digit number (Goyvaerts and Levithan, 2012).

- **Fingerprinting:** Detects data using fingerprinting or indexing structured data sources such as databases. Exact data matching reduces the number of false positives but a good datasource is required. Since the fingerprinting can consume a lot of resources it doesn't usually run on endpoints depending on the size of the data set (Mogull, 2014). Additionally, Costante et al. (2016) presented an approach to detect possible leaks, identifying anomalies in database transactions. They refer to this solution as "white-box", because it creates self-explanatory profiles that are easy to understand and update as opposite to black-box systems which create profiles hard to interpret and maintain (for example, neural networks). With this approach, it is demonstrated: (i) significantly reduces the number of false positives; (ii) creates profiles that are easy to understand and update and therefore provides an explanation of the origins of an anomaly; (iii) it allows the introduction of a feedback mechanism that allows the system to improve from its own errors; and (iv) resource aggregation and transaction flow analysis allow the system to detect threats spanning multiple resources and transactions. Furthermore, Shapira, Shapira and Shabtai (2013) had also investigated fingerprinting. Protecting confidential information from unauthorized disclosure is a major concern for all organizations. Because an organization's employees need access to information to perform their daily work, data leak detection is an essential and challenging task. Fingerprinting is a content-based method used to detect data leakage. In fingerprinting, signatures of known confidential content are extracted and combined with the output content to detect the leak. Existing fingerprinting methods, however, suffer from two main limitations. First, the fingerprint can be circumvented by reformulating (or minor modifying) the confidential content, and second, generally all content in the document is fingerprinted (including non-confidential parts), resulting in false positives. In their work, Shapira et al. (2013) propose an extension of the fingerprinting approach based on ordered k-skip-n-gram. The proposed method is capable of producing a fingerprint of the central confidential content that ignores non-relevant (non-confidential) sections. In addition, the proposed fingerprint method is more robust for redesign and can also be used to detect a confidential document not previously seen and therefore provide better detection of intentional leak incidents.

- **Binary file matching**: This technique creates a hash of the binary file. It may be prone to false positives as a minor change in the file will result in a different hash value.
- **Machine Learning:** Hart, Manadhata and Johnson (2011) presented algorithms for automatic text classification to classify business documents as sensitive or non-sensitive. They also introduce a new training strategy, complement and adjust, to create a classifier that has a low false discovery rate, even when presented with documents not related to the company. The algorithm had a false negative rate of less than 3.0% for all tests (that is, in a real deployment, the classifier can identify more than 97% of information leaks). Moreover, Katz, Elovici and Shapira (2014) address a method called CoBAn. This new method consists of two phases: training and detection. During the training phase, clusters of documents are generated and a graphical representation of the confidential content of each cluster is also created. This representation consists of key terms and the context in which they need to appear to be considered confidential. During the detection phase, each document tested is assigned to several clusters and its content is then combined with the respective graph of each cluster in an attempt to determine the document's confidentiality. They concluded that the model is superior to other methods in detecting escape attempts, where confidential information is reformulated or is different from the original examples provided in the learning set.
- **Lexicon**: Uses a combination of rules, dictionaries and other analysis to protect information.
- **Categories**: DLP Solutions contain by default pre-defined templates that can be reused for common types of data, such as PCI, HIPAA, PII, among many others.

2.7.Data Loss Prevention Solutions

This section will focus on doing research in the different Data Loss Prevention solutions available and key capabilities. This will allow to understand the options available when choosing a solution and the advantages and disadvantages of each one. Data Loss Prevention solutions can be distinguished in two categories:

- **Enterprise DLP**: Incorporates sophisticated detection techniques to allow organizations address their data protection concerns. Products are packaged with physical and virtual appliances for management, agents or data discovery. Leading characteristics of Enterprise DLP includes a centralized management console, advanced policy definition, event management, workflow and reporting. DLP system works as a centralized system for data protection within an organization to mitigate the risk of data loss at the endpoint, in storage and over the network (Reed and Wynne, 2017).
- **Integrated DLP**: Integrated DLP offer a more limited set of functionalities that typically enforce policies on one specific type of data, usually, data in motion, over a specific channel (e.g. email). Integrated DLP are content-aware, but most focus on a set of regulatory compliance and basic intellectual property use cases (Reed and Wynne, 2017; Radicati, 2017).

There are a number of commercial DLP solutions available:

Figure 8 - 2017 Gartner Magic Quadrant for Enterprise Data Loss Prevention

Figure 8 source: Reed and Wynne, 2017.

This thesis will focus on the comparison of the solutions of the "Leaders" as per the figure above. While doing the research, it was concluded that open-source data loss prevention solutions have been acquired by different vendors, however, OpenDLP remains available in version 0.5.1 (released in August 2012). OpenDLP will also be

compared with the commercial DLP solutions. The investigation about the different DLP solution consisted in analyzing product documentation, reading technical papers available for the solution including market research company's information.

2.7.1. Digital Guardian

Digital Guardian provides a platform to stop data loss from insiders and malicious data theft from outside attacks. Endpoint capability also includes endpoint detection and response (EDR). The platform covers endpoint, network, cloud applications and reporting. Main components of Digital Guardian solution are:

- Protection Platform: The platform can be deployed in different ways: on-premises, SaaS or managed service. The purpose of the platform is to discover and protect sensitive information across the organization. It also covers endpoint, network, cloud and mobile devices using context-based and content based (fingerprinting) classification.
- Endpoint: The endpoint module provides user-based classification of sensitive data, analyses the content of data with DLP engine and enforces the policies across all egress channels. The agent is available on Windows, macOS and Linux.
- Network: The network module provides protection and discovery for sensitive data. The module prevents information leakage in multiple communication channels including email, web, FTP and SSL.
- Cloud: This module integrates with cloud storage and collaboration providers such as Box, Citrix and Microsoft. It gives visibility of sensitive data stored in cloud storage and monitors files that have been uploaded. Depending on the policies configured, automatic remediation is possible, as well as, alerting administrators and data owners of the incident.
- Analytics and Reporting: This module is an advanced report and analytics solution in the form of a cloud solution. It leverages the endpoint agent and network appliances to provide in-depth visibility of the system, user and events. The same console is also leveraged for endpoint detection and response.

2.7.2. Forcepoint

Forcepoint DLP provides four different solutions that deployed together form the overall Forcepoint DLP.

DLP Endpoint: Protects data on endpoints supporting Windows, Linux and macOS. The endpoint module addresses data in motion, data in use and data at rest use cases.

DLP Cloud Applications: This module protects data in cloud applications such as Office 365, G Suite, Box and others. This module is provided through the integration with Forcepoint CASB.

DLP Network: This module monitors sensitive data sent outside the organization applying the DLP polices configured. It can apply different actions such as alert, block, notify and quarantine data in email, web, FTP, among different channels. It also supports integrated OCR for different languages.

DLP Discover: This module performs data discovery by using an agent or agentless in fileservers, databases, collaboration platforms and email servers (both on-premises and in the cloud). The identified information can then be encrypted, removed, quarantined, audited, among other actions. It also supports OCR.

2.7.3. Intel Security

Intel Security provides a number of different DLP components that builds the overall solution. The main components are as follows:

Device Control: This module monitors the copying of sensitive data to external media such as storage devices, CD, DVD, Bluetooth, among others.

DLP Discover: This module focuses on data at rest use-cases and identifies and protects sensitive information. It supports a third-party cloud storage solution – Box, and the solution can also identify data in the network, including databases and sharepoint.

DLP Monitor: This module focuses on data in motion use-case and identifies, monitors and tracks information in the organization.

DLP Prevent: This module encrypts, quarantine or block sensitive information being sent over email, instant messaging, HTTP/HTTPS, FTP, among others. It includes DLP

Prevent for Mobile Email which intercepts emails downloaded to mobile devices using ActiveSync.

DLP Endpoint: This module monitors sensitive information on endpoints, but also support information that is being copied to the network, removable storage devices and others. The solution can also block, alert, notify, encrypt, quarantine and perform different actions on sensitive information on an endpoint.

2.7.4. OpenDLP

OpenDLP is a free and open source DLP solution, developed by Andrew Gavin that offers a central management and an agent. OpenDLP has two components:

Web Application: This module acts as the management and allows to deploy agents over SMB, create reusable policies for scans, review the findings and identify any false-positive. It also allows to create regular expressions in order to find sensitive data for data at rest use-cases. The Web Application is written in Perl and uses MySQL backend.

Agent: The agent runs on Windows systems and installs itself as a service, it is written in C and therefore comes with no .NET Framework dependencies. The agent whitelists and blacklists files and directories and pushes the findings to the web application over a secure channel (two-way trusted SSL connection).

In addition to performing sensitive data discovery on Windows, OpenDLP also supports performing agentless data discovery against Microsoft SQL Server and MySQL.

2.7.5. Symantec

Symantec Data Loss Prevention solutions can be consumed in different ways: cloud service, software and virtual or hardware appliances. The Symantec DLP solution contains the following modules.

DLP Sensitive Image Recognition: This module can detect sensitive images and text embedded in images such as scanned documents, screenshots, pictures and PDFs by using a form recognition technology and a built-in OCR.

DLP Cloud Services: This module includes integration with the detection server that monitor sensitive data stored in cloud applications such as Box, Dropbox, Office 365,

among others. It also integrates with CloudSOC CASB and web security services. It also provides integration with email (both on-premises and cloud).

DLP for Network: This module includes different components and monitors and inspects sensitive information across the network. It includes the network monitor that analyses data across different protocols, such as, SMTP, HTTP/HTTPS, including custom protocols and also IPv6. The prevent for email integrates with corporate email to detect and block information leakage over this channel. It supports by using network prevent for web, the monitoring of web traffic.

DLP for Endpoint: This module can discover and block sensitive data stored on an endpoint leaving to different channels (removable media, printing, copy/move to network shares, among others).

DLP for Storage: This module scans and secures sensitive data stored on different repositories, including, NAS, databases, exchange and sharepoint servers, among others.

Information Centric Tagging and Encryption: The tagging module allows the ability to apply tags and watermarks to classify sensitive information; It can automatically apply a tag based on a DLP policy. The encryption component allows to apply digital rights to a document as a response of a DLP policy. The encryption follows the file and can be remotely deleted.

2.7.6. DLP Solutions Comparison

As a summary for this analysis, the commercial vendors offer a complete set of modules and components that allows organizations to reduce the risk of data loss. Every commercial vendor covers data in use, data at rest and data in motion use-cases with discovery and protection capabilities that scales to cloud platforms as well. OpenDLP is no longer maintained and updated and therefore it is not considered as a solid alternative to commercial products.

The table below contains a list of the offerings and the main components available.

Table 2 - DLP Product Comparison

Feature / Vendor	Digital Guardian	Forcepoint	Intel Security	OpenDLP	Symantec
Data State					
Data at Rest	Yes	Yes	Yes	Yes	Yes
Data in Use	Yes	Yes	Yes	Yes	Yes
Data in Motion	Yes	Yes	Yes	No	Yes
Information Classification					
Can classify	No	Partial [1]	Yes	No	Yes
Remediation					
Audit	Yes	Yes	Yes	Yes	Yes
Block/Remove	Yes	Yes	Yes	No	Yes
Notify	Yes	Yes	Yes	No	Yes
Encrypt	Yes	Yes	Yes	No	Yes
Quarantine	No	Yes	Yes	No	Yes
Miscellaneous					
Endpoint support (Win, MAC, Linux)	Yes	Yes	Partial	Partial	Partial
Common Policy[2]	No	Yes	Yes	No	Yes
CASB Functionality	No	Yes	Yes	No	Yes

[1] Integrates with third-party.

[2] Common policy allows to apply a single policy to all data states.

2.8.Conclusion

This chapter covered the traditional security solutions implemented by organizations and a literature review around data breaches and data loss prevention. The latter, explaining how this solution work, both from a data state perspective but also from a feature capability such as detection techniques. A feature comparison matrix was built to compare and highlight differences between different investigated DLP solutions.

Chapter 3 – Proposed Solution and Implementation

3.1.Introduction

Based on the investigation around multiple DLP Solutions available, the proposed architecture will cover the three main data loss vectors: Data in Motion, Data at Rest and Data in Use. Although each component will address specific needs, working in conjunction they will effectively decrease the risk of data loss.

Based on the research, the proposed architecture takes into consideration a lack of consistency between the different DLP solutions available from an architecture perspective since each solution presents a different architecture and modules that can be misleading. The proposed architecture discussed in this chapter was chosen as it focuses on protecting the information in different states, adapts to different scenarios and identifies the main components that should be implemented to properly identify and protect confidential information leaving the organization or unauthorized locations. Moreover, the architecture is flexible to adapt to different use cases and addresses the following requirements:

- Identify confidential information across the organization in different vectors: Data in Motion, Data at Rest and Data in Use.
- Allows the development of new DLP Policies: Managing policies is the core capability of a DLP solution. A centralized management console should allow to properly author new policies using different techniques in order to identify confidential information.
- Allows the integration with third-party solutions: Integration with active directory is common to allow administrators to login into the management console and also to get more context of the users that can be shown in a DLP incident. Export of incidents and logs information should also be possible.
- Provide visibility into DLP incidents: Once a policy is triggered, it should be possible to gather all the information regarding the incident.
- Reporting: Provides the capability to access reports and generate new reports.

3.2.Proposed Solution High Level Design

The following diagram represents the high-level architecture of the proposed data loss prevention solution:

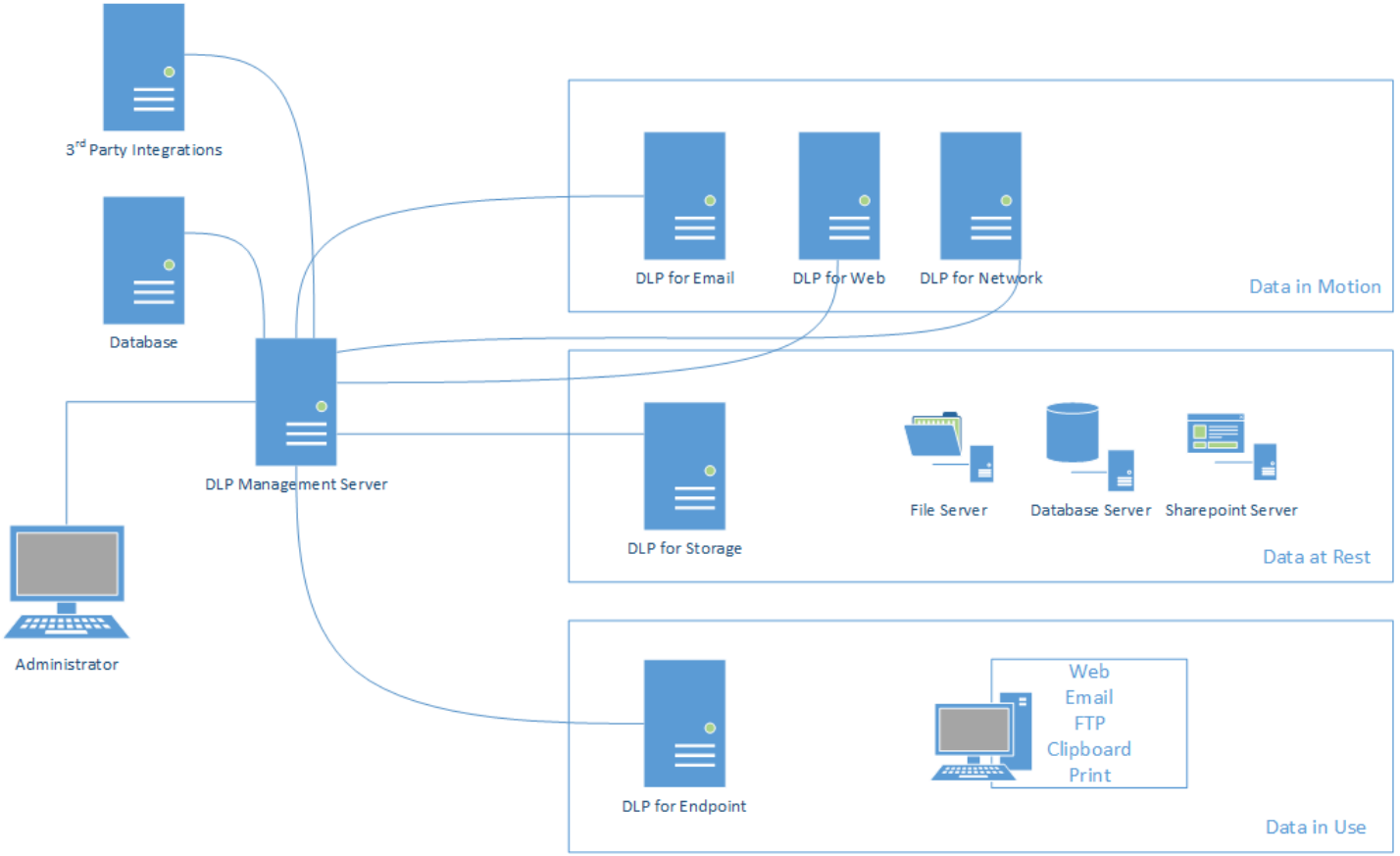

Figure 9 - DLP High Level Design

3.2.1 High Level Design Components

This section describes the components of the proposed architecture.

- Database Server: This server will host the database for the DLP Solution. Depending on the solution to be implemented the database software may differ.
- DLP Management Server: This is the central management server where policies are created and pushed to other DLP servers, incidents and workflows are managed. Reporting is also a component that sits in the management server. This is also the server where third-party solutions can be integrated – some examples include Active Directory to retrieve attribute information about users generating incidents (name, location, email, etc.) and Security Information and Event Management (SIEM) to centrally store incident logging for further analysis.
- DLP for Endpoint: This server will protect the endpoints and have different roles: It can scan information stored locally and also monitor multiple channels, such as, copying files to network shares, email, removable media, printing, HTTP and cloud applications. It also can block, alert, notify, encrypt, quarantine, and perform different actions on an endpoint.
- DLP for Storage: This component has the capability of detecting confidential information saved across a variety of sources, such as endpoint devices, file servers, websites, web portals and databases. Once the confidential information is identified, policies can be applied to create new security incidents, quarantine or encrypt the information, essentially remediating the information identified.
- DLP for Email: The email component will integrate with existing MTA to analyze the contents of the emails; it can analyze email headers, body and any attachment and apply a remediation policy, if needed. Typically, the email component of the DLP Solution will receive the email from an upstream MTA, analyze its contents and forward it to a downstream MTA, however, it is also

possible to use a single MTA. The following diagrams explains the email flow when using a single MTA:

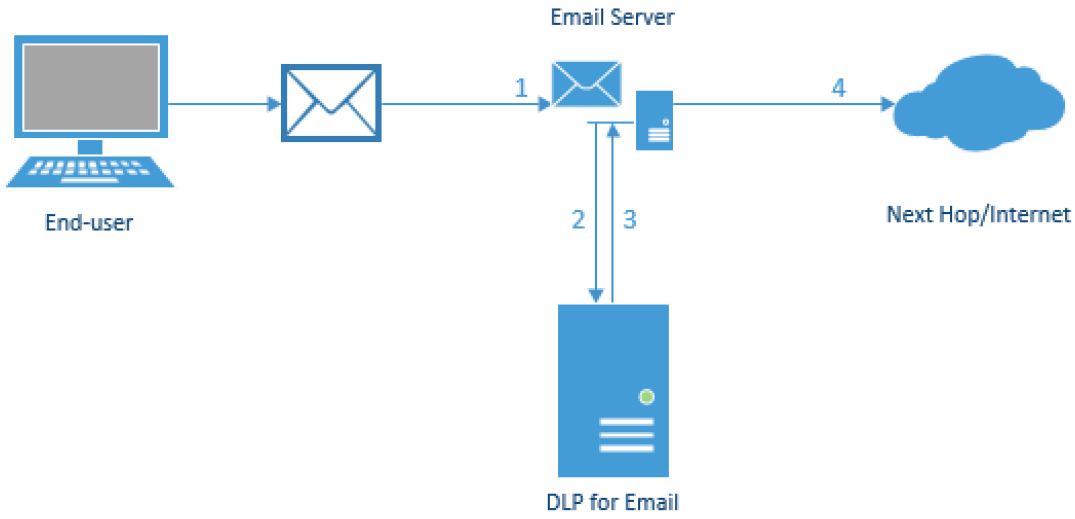

Figure 10 - DLP for Email Single MTA Architecture

1. End-user sends and email that arrives to the email server
2. Email server forwards the email to the DLP for Email server
3. DLP for Email analysis the contents of the email and determines whether a policy is matched. Content can be blocker and email headers can be added so the email server knows the email was already processed by DLP for Email
4. If DLP for Email does not match any policy, email is sent to its intended recipient

When the DLP for Email uses two MTAs, the architecture and email flow is as follows:

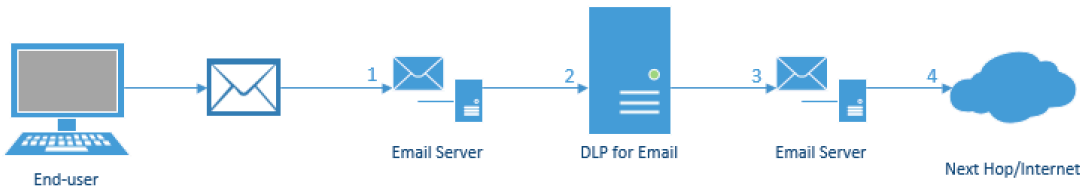

Figure 11 - DLP for Email using multiple MTA Architecture

1. End-user sends and email that arrives to the email server

2. DLP for Email analysis the contents of the email and determines whether a policy is matched. Any action such as block the email or notify the sender can be taken
3. If DLP for Email determines that the email is safe to be sent it will send it to the next email server or MTA
4. Email server or MTA sends the email to the intended recipient

- DLP for Web: With the web component it is possible to intercept web communications (HTTP/HTTPS). Confidential information leaving the company over this channel can be monitored and potentially, blocked; integration between the web proxy and the DLP for Web server is made using Internet Content Adaptation Protocol (ICAP) protocol. The following diagram represents the web traffic flow when integrating with DLP for Web:

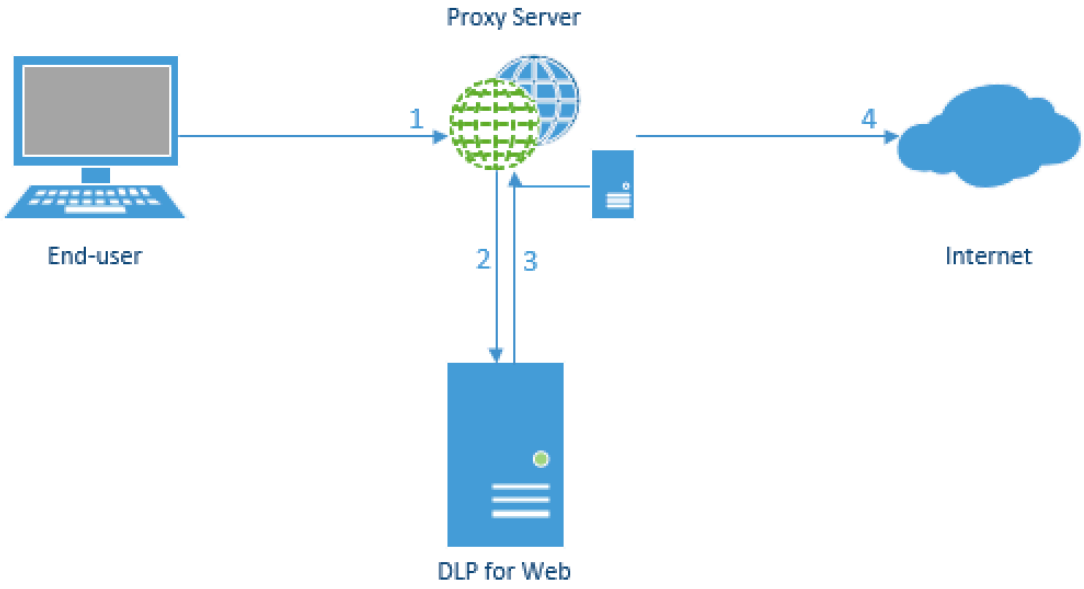

Figure 12 - DLP for Web Architecture

1. End-user browses the internet and requests arrive to the proxy server
2. The proxy server (ICAP Client) is integrated with the DLP for Web server (ICAP Server) and forwards to traffic
3. DLP for Web analysis the web traffic and sends a response back to the proxy server to either allow or block the web traffic
4. If DLP for Web does not match a policy and allows the traffic, it is sent to its destination

- DLP for Network: This component allows to receive a copy of the packets either through a port-span or a network tap. This is mainly a monitoring only component as it is not inline with the traffic. However, receiving a copy of the traffic will allow to identify confidential information in the network segment on which the span is configured. The following diagram represents the monitoring of traffic using a port-span:

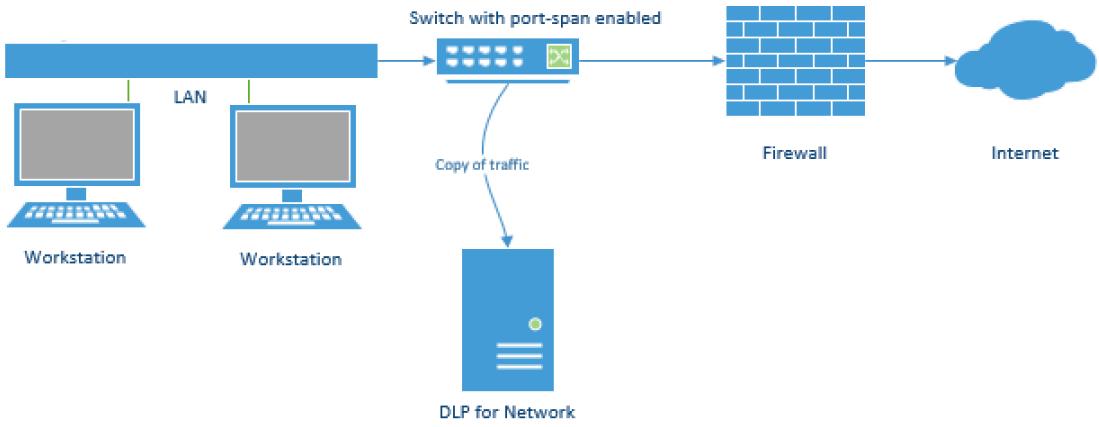

Figure 13 - DLP for Network Architecture

In this architecture, a network segment is connected to a switch that has a port-span configured; all traffic that transverses this switch will be copied over to the DLP for Network. Since DLP will only receive a copy of the traffic, blocking is not possible however, in case a DLP Policy is triggered it will be logged in the DLP Management Console.

3.3.DLP Implementation

For the practical implementation of the proposed solution architecture, it will be demonstrated using Symantec Data Loss Prevention. The reason of using this solution is that, no open-source solution is mature enough to present more advanced DLP use-cases and this solution is considered, as per market research companies such as Gartner, the leader in Data Loss Prevention.

Although this DLP solution contains different modules and components, the implementation will focus on Data in Use, Data at Rest and Data in Motion as described in the proposed architecture.

This DLP Solution offers a rich feature set and lets organizations safeguard data, company information, intellectual property, and sensitive or classified information, whether it is: exiting the network via corporate email, web mail, or other Internet protocols (DLP for Network); exiting endpoints via USB, CD/DVD, network protocols, and so on, or stored on endpoints (DLP for Endpoint); stored on shared servers and data repositories (DLP for Storage); or exiting mobile devices via corporate email (Exchange ActiveSync), web mail, web posts, or mobile app (DLP for Mobile); or via email sent through Microsoft Office 365 cloud services (DLP for Cloud).

This solutions consists of the Enforce Platform management console and several associated modules: DLP Network Monitor and Network Prevent for Email and Web (DLP for Network); DLP Endpoint Discover and Endpoint Prevent (DLP for Endpoint); DLP Network Discover, Network Protect, Data Insight (DI) and Data Insight Self-Service Portal (DLP for Storage), and DLP Mobile Email Monitor, Mobile Prevent and Cloud Prevent for Office 365 (DLP for Mobile/Cloud).

3.3.1 Proposed Architecture Mapped to Symantec Data Loss Prevention

Symantec Data Loss Prevention modules and components can be easily adapted to the proposed generic Data Loss Prevention:

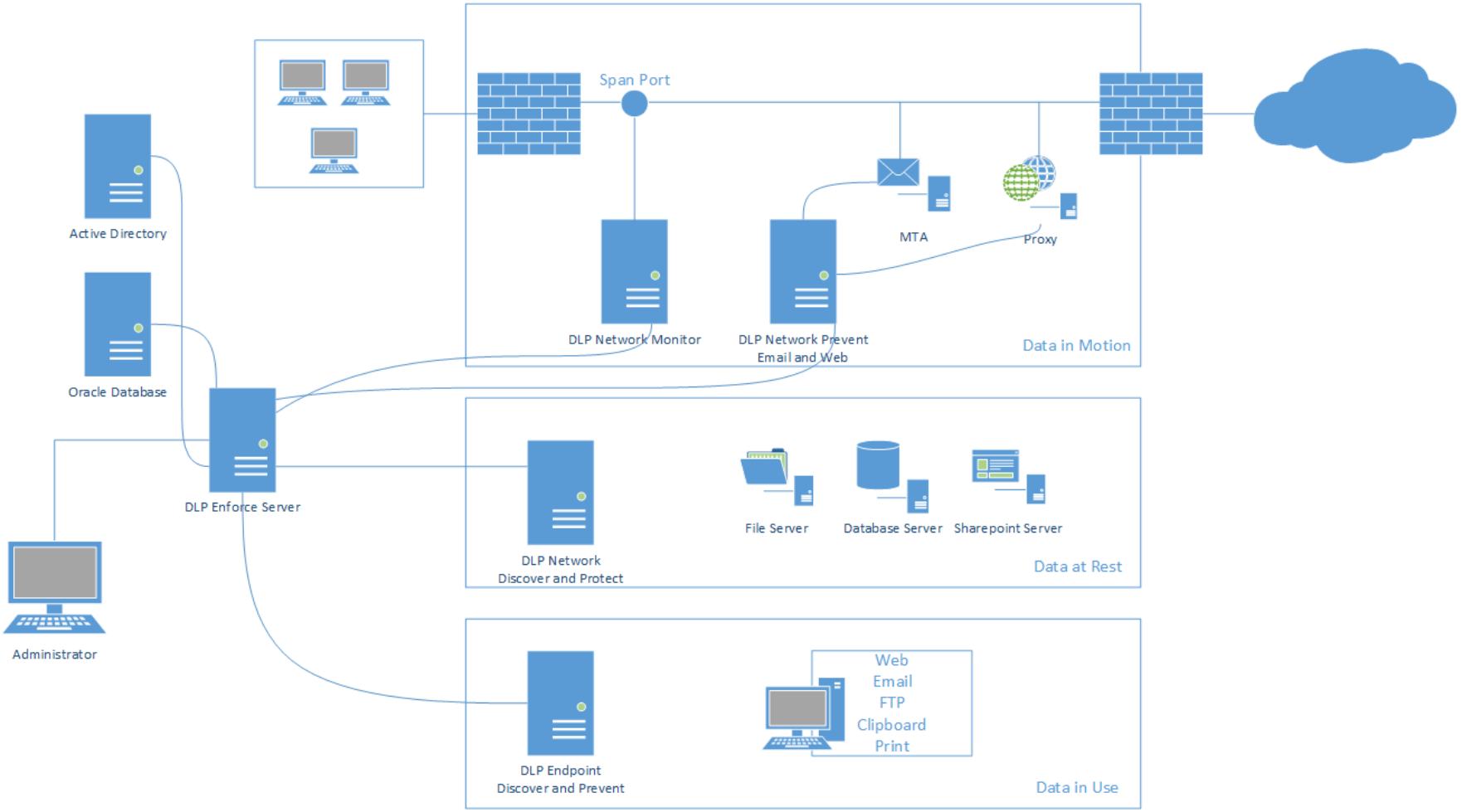

Figure 14 – Proposed Architecture Mapped to Symantec Data Loss Prevention

With this architecture, we cover Data in Use with the deployment of a Symantec DLP Endpoint Discover and Protect (components responsible to discover and protect information at the endpoint), Data at Rest is assured by the use of Symantec DLP Network Discover and Protect (components that allow to scan information stored in filesystems or other storage devices) and for Data in Motion, Symantec provides a DLP Network Monitor to integrate with the network using a port-span or network tap and DLP Network Prevent for email and web to integrate with email gateways and web proxies. The Symantec DLP components were also described in section 3.3. Each one of these DLP servers are also known as detection servers.

3.3.2 Installation Tiers

Symantec Data Loss Prevention can be implemented in different ways. The types of implementations are known as Installation Tiers.

There are three installation tiers:

- Single Tier: DLP Enforce Server (Management Server), Oracle Database and any Detection Server. Single tier is applicable only for lab and test environments.

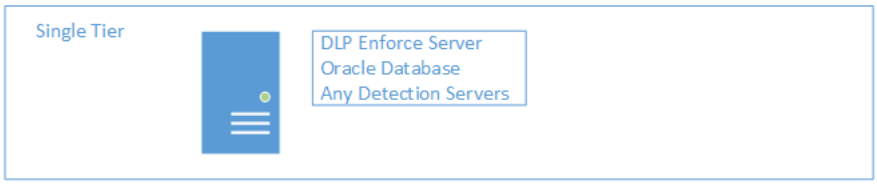

Figure 15 – Single tier DLP deployment

- Two Tier: DLP Enforce Server and Oracle Database on the same server. Detection Servers installed on dedicated servers.

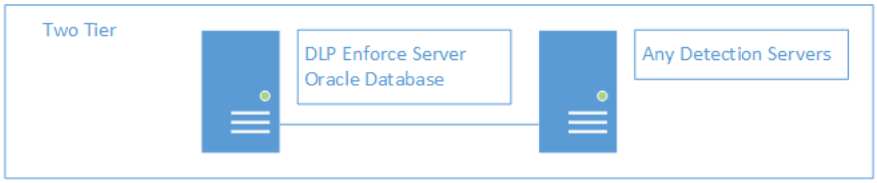

Figure 16 – Two-tier DLP deployment

- Three Tier: DLP Enforce Server, Oracle Database and Detection Servers installed on dedicated Servers.

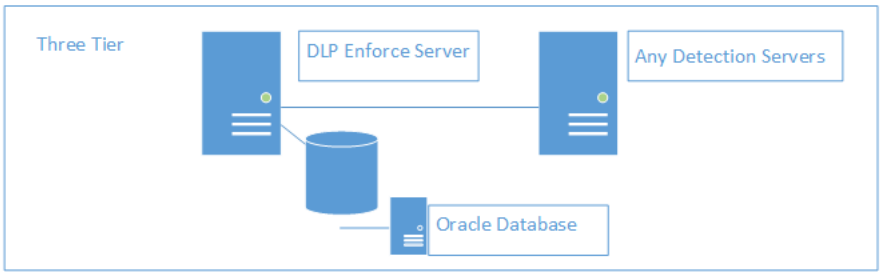

Figure 17 – Three-tier DLP Deployment

For the implementation to be used in this thesis it will be used the Single-Tier installation.

3.3.3 System Requirements for Test Environment

The Test environment uses three virtual machines. The first virtual machine is the DLP Single Tier Server that hosts the Enforce Server, Oracle Database and Detection Servers. The second server hosts the Active Directory and other services such as DNS, DHCP and File Server Roles. The third virtual machine represents a user workstation running Windows 10. Below are the specifications for each virtual machine.

This environment will be used to support testing and validation of the architecture in chapter 4.

DLP Single Tier			
CPU	RAM	Disk	Operating System
8 cores	64GB	3TB	Windows Server 2016

Active Directory File Server

CPU	RAM	Disk	Operating System
4 cores	16GB	500GB	Windows Server 2016

End User Machine			
CPU	RAM	Disk	Operating System
1	4GB	256GB	Windows 10

The installation of the DLP Single Tier Server can be found in Appendix A.

3.3.4 Solution Pack

Symantec Data Loss Prevention uses the concept of Solution Packs which contains configured DLP policies, response rules (block, notify, quarantine), user roles, reports, protocols, and the incident statuses (new incident, closed, under investigation, etc.) that support a particular industry or organization. A solution pack essentially provides a baseline of configurations instead of being required the configuration of everything from scratch. Importing a solution pack right after the initial Enforce Server installation provides a baseline of configurations that can be used to configure the platform. A solution pack is specific to Symantec DLP and the policies or different configurations cannot be exported to different DLP solutions and must be manually replicated. Each DLP solution have its own way to have a baseline of configurations but all of them allow to manually create and configure the solution.

The Solutions Packs available are:

Table 3 - Solution Packs

Name	Filename

Data Classification for Enterprise Vault Solution Pack	Data_Classification_Enterprise_Vault_v14.6.vsp
Energy & Utilities Solution Pack	Energy_v14.6.vsp
EU and UK Solution Pack	EU_UK_v14.6.vsp
Federal Solution Pack	Federal_v14.6.vsp
Financial Services	Financial_v14.6.vsp
Health Care Solution Pack	Health_Care_v14.6.vsp
High Tech Solution Pack	High_Tech_v14.6.vsp
Insurance Solution Pack	Insurance_v14.6.vsp
Manufacturing Solution Pack	Manufacturing_v14.6.vsp
Media & Entertainment Solution Pack	Media_Entertainment_v14.6.vsp
Pharmaceutical Solution Pack	Pharmaceutical_v14.6.vsp
Retail Solution Pack	Retail_v14.6.vsp
Telecom Solution Pack	Telecom_v14.6.vsp
General Solution Pack	Vontu_Classic_v14.6.vsp

During the implementation of the DLP Solution, it will be imported the Financial Services solution pack. This will provide a solid baseline for policies templates as well as key configurations. The configuration of the solution pack is also described in Appendix A.

3.3.5 DLP Agent Configuration and Installation

The Symantec Endpoint agent software reserves a minimum of 30 MB of memory on the Endpoint computer. The Endpoint agent software temporarily consumes additional

memory while it detects content or communicates with the Endpoint Prevent server. After these tasks are complete, the memory usage returns to the previous minimum. System utilization can also be throttled based upon a variety of factors.

The initial endpoint agent installation consumes approximately 70 MB to 80 MB of hard disk space. The actual minimum amount depends on the size and number of policies that are deployed to the endpoint computer. Additional disk space is then required to temporarily store incident data on the endpoint computer until the endpoint agent sends that data to the Endpoint Prevent server. If the endpoint computer cannot connect to the Endpoint Prevent server for an extended period of time, the endpoint agent will continue to consume additional disk space as new incidents are created. The disk space is freed only after the agent software reconnects to the Endpoint Prevent server and transfers the stored incidents. The amount of disk-space utilized for incident-storage can also be modified.

All data stored by DLP at the Endpoint is protected from user access and manipulation. All communications are encrypted and require matching keys to ensure that no Endpoint can be redirected to a rogue Endpoint-Server. The DLP Agent can be configured to connect to multiple Endpoint Servers. Multiple Endpoint Servers enable incidents and events to be sent to the Enforce Server in a timely way if an Endpoint Server becomes unavailable. The DLP Agent, after a specified amount of time, connects to another Endpoint Server to transmit the incidents and events that it has stored. The DLP Agent fails over to a different Endpoint Server only when the current Endpoint Server is unavailable. If the original Endpoint Server is unavailable, the Agent attempts to connect to another Endpoint Server in the configured list. By default, the DLP Agent tries to reconnect to the original Endpoint Server for 60 minutes before it connects to another Endpoint Server.

When the DLP Agent connects to a new Endpoint Server, it downloads the policies from that Endpoint Server. It then immediately begins to apply the new policies. To ensure consistent incident detection after a failover, maintain the same policies on all Endpoint Servers to which the DLP Agent may connect.

For Endpoint Discover monitoring the DLP Agent downloads the new Endpoint Discover scan configuration and policies from the new Endpoint Server and immediately runs a new scan. The new scan runs only if there is an active Endpoint Discover scan configured on the new Endpoint Server.

While configuring the deployment of DLP Agents, the detection server lists needs to be configured as well. The procedure for adding a list of Endpoint Servers appears under each method of installation. An IP addresses or host names with the associated port numbers can be specified. If a host name is used, then the DLP Agent performs a DNS lookup to get a set of IP addresses. It then connects to each IP address. Using host names and DNS lookup allows to make dynamic configuration changes instead of relying on a static list of stated IP addresses.

For the installation and configuration of the agents, a new configuration was performed:

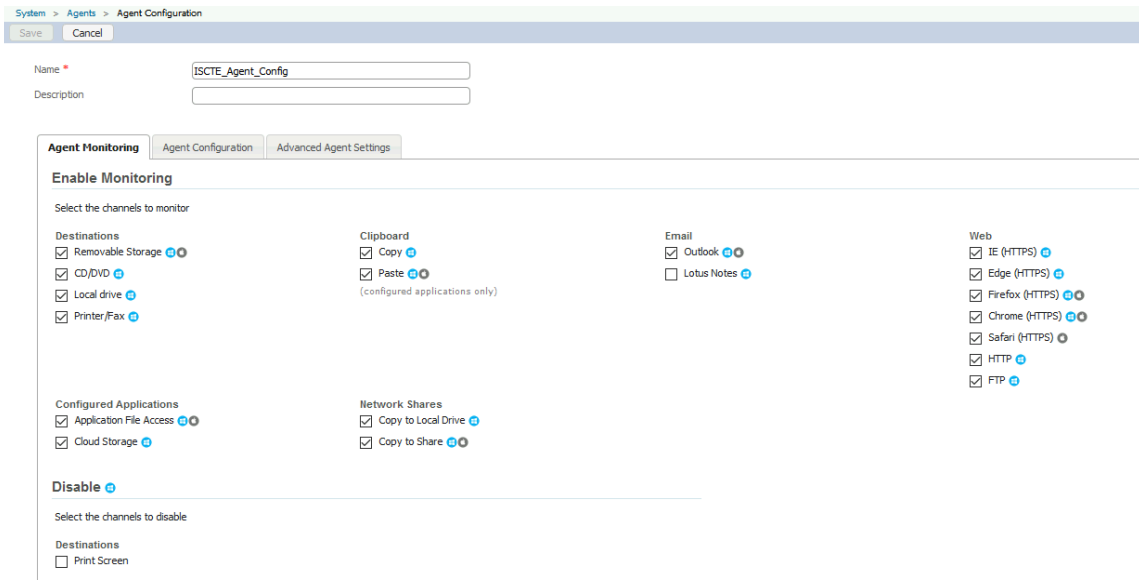

Figure 18 - DLP Agent Configuration

The default settings were accepted, except for the Agent Monitoring tab that the monitoring was configured for:

- Destinations: Monitors copying of information to removable storage, CD/DVD, local drive and printer/fax.
- Clipboard: Monitors the clipboard for copy and past actions
- Email: Monitors emails that uses Outlook as the email client
- Web: HTTPs monitoring was enabled at the endpoint level for Internet Explorer, Edge, Firefox, Chrome and Safari as well as communications using HTTP and FTP protocol
- Configured Applications: Monitors supported application file access (data leaving applications on endpoints) and cloud storage.

- Network Share: Monitors copies to local drive and to network shares.

After the DLP Agent configuration the packaging as configured:

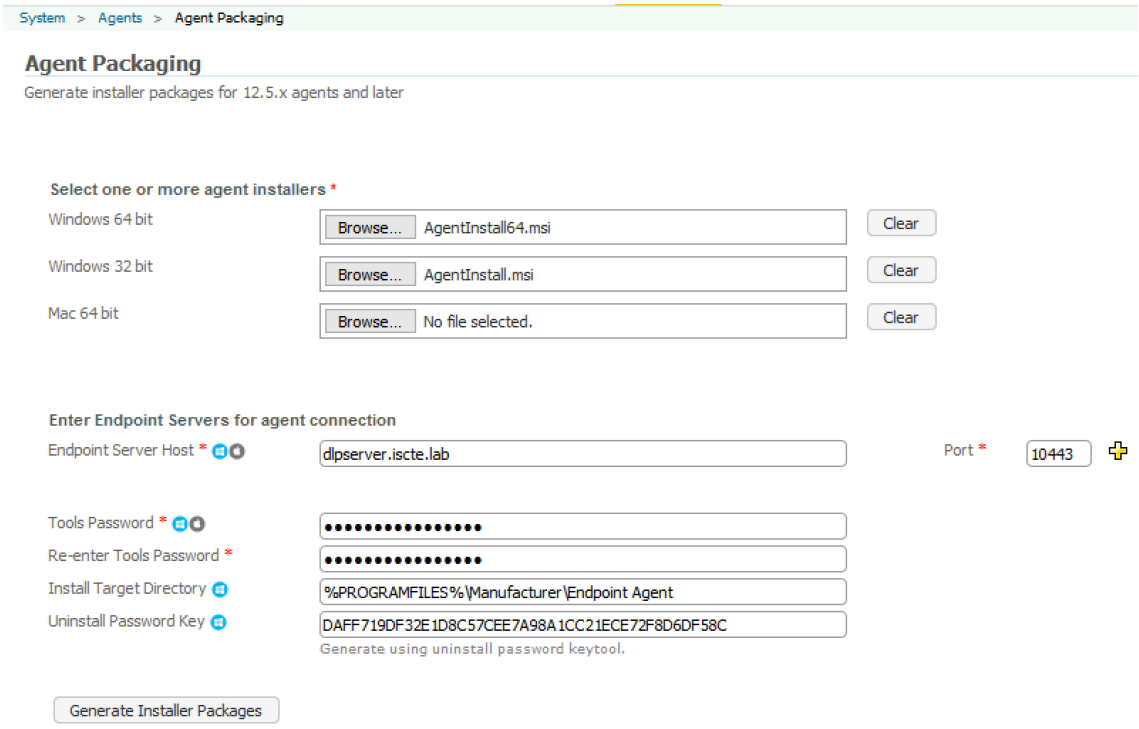

Figure 19 - DLP Agent Packaging

After clicking the Generate Installer Packages, two different installation packages are generated: for Windows 64 bit and Windows 32 bit.

The table below contains the files generated for installation.

Table 4 - DLP Agent Installation Files

Filename	Description
AgentInstall.msi (or AgentInstall64.msi)	MSI installation file
endpoint_cert.pem	Endpoint self-signed certificate
endpoint_priv.pem	Endpoint private key
endpoint_truststore.pem	Endpoint trust store to trust the server public key
install_agent.bat	Batch file used for installation

PackageGenerationManifest.mf	Package metadata
upgrade_agent.bat	Batch file used for agent upgrade

To perform the installation the install_agent.bat file is used. Below is the content of the installation file:

```
msiexec /i AgentInstall.msi /q
INSTALLDIR="%PROGRAMFILES%\Manufacturer\Endpoint Agent"
ENDPOINTSERVER="dlpserver.iscte.lab:10443"
TOOLS_KEY="DC27DB3443DC819BB3EB2989832712D1FC46B3DCBF27283A009AE7366707CC
A8A06E3D0D8EF27B82E3661EAD6B1223CD04CBFEDA5F07CD42A909F79D5D63BBCD34A3702
305A53"
UNINSTALLPASSWORDKEY="DAFF719DF32E1D8C57CEE7A98A1CC21ECE72F8D6DF58C"
SERVICENAME="EDPA" WATCHDOGNAME="WDP" ARPSYSTEMCOMPONENT="1"
ENDPOINT_CERTIFICATE="endpoint_cert.pem"
ENDPOINT_PRIVATEKEY="endpoint_priv.pem"
ENDPOINT_PRIVATEKEY_PASSWORD="7F8A1B59F55BC171FE7A09C1DC76CEE8319AF13FD1A
F006600811C9F4071E64BB738DCB04863F45B133138DF634725F5C97B6C57013DC91B358A
CFC6142701874553C2C5BA9AA38EB48074B8ACD7AD1E421F93133E263FE1963E032F76911
2ECC5C81" ENDPOINT_TRUSTSTORE="endpoint_truststore.pem" LOGDETAILS="Yes"
/L*v %SystemDrive%\installAgent.log
```

The batch file contains

Table 5 - DLP Agent Command-Line Arguments

Command/Argument	Description
msiexec	The Windows command for executing MSI packages.
/i	Name of the installation package
/q	Specifies that a silent install should be performed
INSTALLDIR	Specifies the installation directory

ENDPOINTSERVER	Specifies the endpoint server and port to which agent connects to
TOOLS_KEY	The password that is associated with the agent tools. This value is defined during the agent installation packaging process.
UNINSTALLPASSWORDKEY	The password the administrator uses when uninstalling agents. This value is defined during the agent
SERVICENAME	The agent service name
WATCHDOGNAME	The watchdog service name
ARPSYSTEMCOMPONENT	Specifies whether the agent software will be displayed in windows “add and remove programs”. A value of 0 will show the software available where a value of 1 will hide
ENDPOINT_CERTIFICATE	The endpoint self-signed certificate file name: endpoint_cert.pem. This file is created during the agent installation packaging process.
ENDPOINT_PRIVATEKEY	The endpoint private key file name: endpoint_priv.pem. This file is created during the agent installation packaging process.
ENDPOINT_PRIVATEKEY_PASSWORD	The password that is associated with the agent certificates. The password is located in the endpoint_priv.pem file,

	which is created during the agent installation packaging process.
ENDPOINT_TRUSTSTORE	The endpoint trust store file to trust the server certificate (server public key): endpoint_truststore.pem. This file is created during the agent installation packaging process.
LOGDETAILS	Specifies whether the installation will save the details, and to which file the logs will be saved

After the DLP Agent is installed, the DLP processes can be seen in the windows task manager:

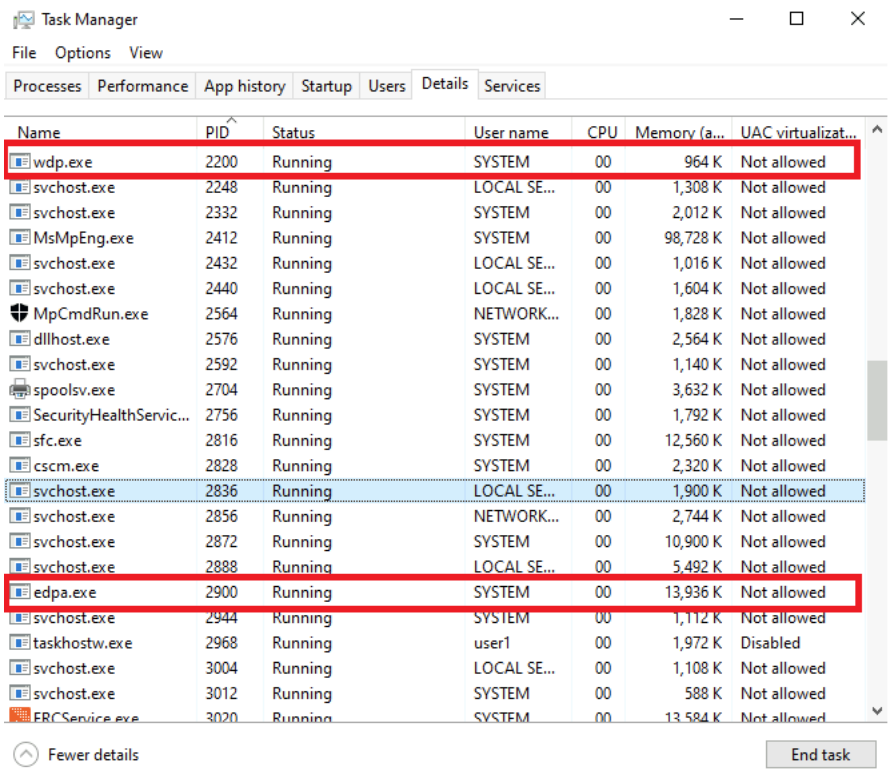

Figure 20 - DLP Agent in Windows Task Manager

3.3.6 Policies

Data Loss Prevention policies are the core of any Data Loss Prevention solution. Policies will detect and prevent data loss; however, policy authoring is just a start of a Data Loss Prevention program, processes are very important to determine what will happen once an incident is generated. If a policy rule is violated, the system creates an incident which can be act on. A single policy can have multiple rules and can be created to target different data loss vectors (data at rest, data in motion, data in use), depending on the policy response, described below, it is possible to apply different actions based on the data loss vector.

Policies within the context of this DLP solution are made of three main components:

- Detection: The detection component is where we specify the detection techniques we will use in this particular policy. Symantec Data Loss Prevention supports multiple detection techniques:
    - The first one is the Exact Data Matching (EDM). EDM allows to detect personally identifiable information (PII), such as social security numbers, bank account numbers, credit card numbers, confidential customer and employee records, and other confidential data stored in a structured data source, like a database, directory server, or a structured data file such as CSV or spreadsheet.
    - Indexed Document Matching (IDM). IDM is used to protect confidential information that is stored as unstructured data in documents and files. For example, you can use IDM to detect financial report data stored in Microsoft Office documents, merger and acquisition information stored in PDF files, and source code stored in text files.
    - Vector Machine Learning (VML). VML performs statistical analysis to protect unstructured data. The analysis determines if content is similar to example content you train against. With VML there is no need to locate and fingerprint all of the data that needs to be protected, instead the system is trained to learn the type of content to be protected based on example documents provided. It is very important to select the correct documents that are representative of the information we want to

protect, also, the opposite applies, we also need to select which files that are not related to the files we want to protect.

- Form Recognition provides the capability of identifying confidential information stored in forms such as tax forms, medical forms, etc. Only specific types of files are supported by Form Recognition. A scoring system exists for this kind of detection technology on which a number between 1-10 identified whether the form is partially filled-in (where 1 is a minimally filled-in form, and 10 is an entirely filled-in form).
- Directory Group Matching (DGM). DGM is used to detect data based on the exact identities of users, senders, and recipients of that data. Using synchronized DGM, you can connect the Enforce Server to a group directory server such as Microsoft Active Directory and detect users based on their directory group.
- Described Content Matching (DCM). DCM is used to detect content and context using different techniques; regular expressions, match specific keywords or even file properties such as filetype, size, among others. DCM is a detection engine very easy to configure and implement, however, it may be prone to false positives if not properly configured. It is recommended to look at compound conditions to proper tune the DCM policy; a compound condition is a rule with multiple conditions declared (A and B), in order for the policy to trigger, both conditions must match.
- Regular Expressions: The DLP solution implements a regular expression engine compatible with Perl Compatible Regular Expressions (PCRE), that provides a mechanism for identifying strings of text, such as particular characters, words, or patterns of characters. Regular expressions can be used to match (or exclude from matching) characters, patterns, and strings.
- Data Identifiers: Data identifiers are algorithms that combine pattern matching with data validators to detect content. Patterns are similar to regular expressions but more efficient because they are tuned to match the data precisely. Validators are accuracy checks that focus the scope of detection and ensure compliance. The solution comes with pre-configured data identifiers that can be used to detect commonly used

sensitive data, such as credit card, social security, and driver's license numbers.

- When creating a detection rule, we also specify a severity for the incident. If we want to detect credit card data, from an incident response perspective, which will be discussed in this section, it may be different the actions we need to take if one credit card is detected, or if hundreds are detected. Severity levels can be defined if the policy rules match a specific match count. If, for example, a match count is less or equal than a ten we can specify the severity as medium, however, it if is higher, than we can set the severity to high.

- Groups: Within group configuration we can configure rules and exceptions. This optional component allows to specify, based on DCM and EDM, to whom will this policy apply. This can be a sender's email, IP Address, windows user, web domain among others.
- Response: The response component of a DLP policy specifies what happens when a policy and rule are triggered. Usually, response rules can be created for the different data loss vectors, and it is possible to block, notify users, modify messages (SMTP), quarantine or copy files, add notes, send incident information to a syslog server and change attributes or statuses of the incidents. Response rules can also target different severity levels. If a detection is made in a low severity incident an incident can be created whether a medium or high severity incident can block or notify different teams for proper escalation. These features will be covered in more detail in chapter 4 with practical examples.

There are other key factors to take into consideration when authoring policies. One is the Policy Groups. When authoring policies, we need to specify a policy group and a policy group is applied to one or more detection servers. It is possible to use policy groups to organize policies and incidents by business units, departments, geographic regions, or any other organizational unit. For example, policy groups for specific departments may be appropriate where security responsibilities are distributed among various groups. With the import of a solution pack, policy groups are already available, they can, however, be edited or changed.

3.4.Conclusion

This chapter discussed the proposed security architecture that mitigates the risk of data loss. It described the need of a generic security architecture that can be adapted to different scenarios and DLP solutions to be deployed by organizations. A DLP solution was chosen and implemented to demonstrate the ease of implementation of the proposed architecture. In addition, the requirements for the implementation of the DLP solution were described including all the necessary steps to successfully deploy the solution.

Chapter 4 – Testing and Validation

4.1.Introduction

This section presents multiple test cases that cover real world scenarios that can be covered with the generic DLP proposed architecture. The aim of these tests is to validate that the proposed architecture mitigates the risk of data loss. It will also ensure that the proposed architecture addresses the use-cases identified and collect evidences that supports the successful detection and prevention of data loss.

4.2.Test Cases

This section will implement a specific number of test cases and collect evidences about the results.

4.2.1. Test Case 1 – PCI DSS Data

Due to compliance reasons, there is the need of identifying credit card information across the network file shares. Information security policy mandates that credit card information can only be stored in a dedicated folder (PCI related data). It is also required from compliance officer that incidents need to be created and properly resolved to safeguard the information. Previous external audits show evidences that credit card data have been found outside of approved locations.

This use case focuses on Data at Rest and requires the following DLP modules to be implemented:

- DLP Management Console (Enforce), including database
- DLP Network Discover and Protect

The first step is to create a DLP Policy that will detect credit card data. The solution already includes a predefined identifier to detect this kind of information.

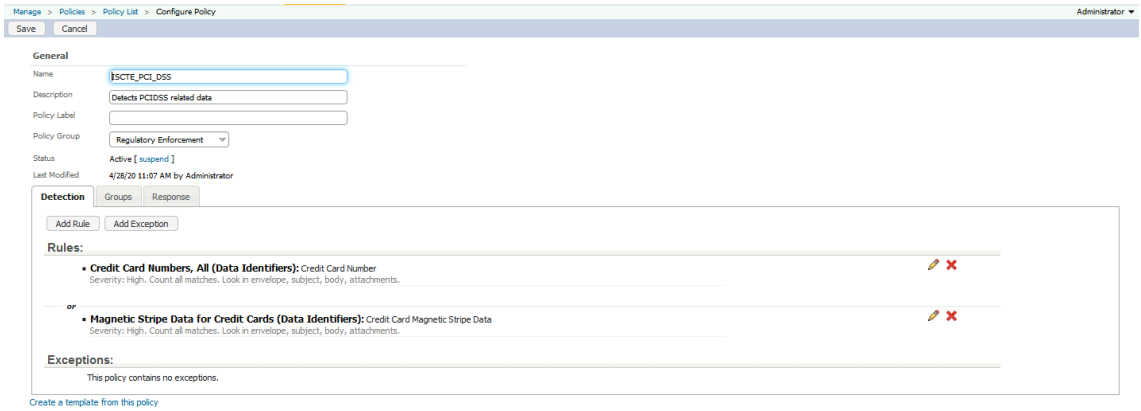

Figure 21 - Policy PCI DSS

This policy was also assigned to a previously created policy group named Regulatory Enforcement so that, from a reporting perspective, we can have information of incidents regarding regulatory policies.

Next, a new Discovery Scanning is configured, it was enabled a weekly scan to be performed in incremental mode (only new or modified items will be scanned in future scans). Pausing capabilities were not enabled, however, if scanning large targets, such as file servers with terabytes of information we could pause scans from running during business hours:

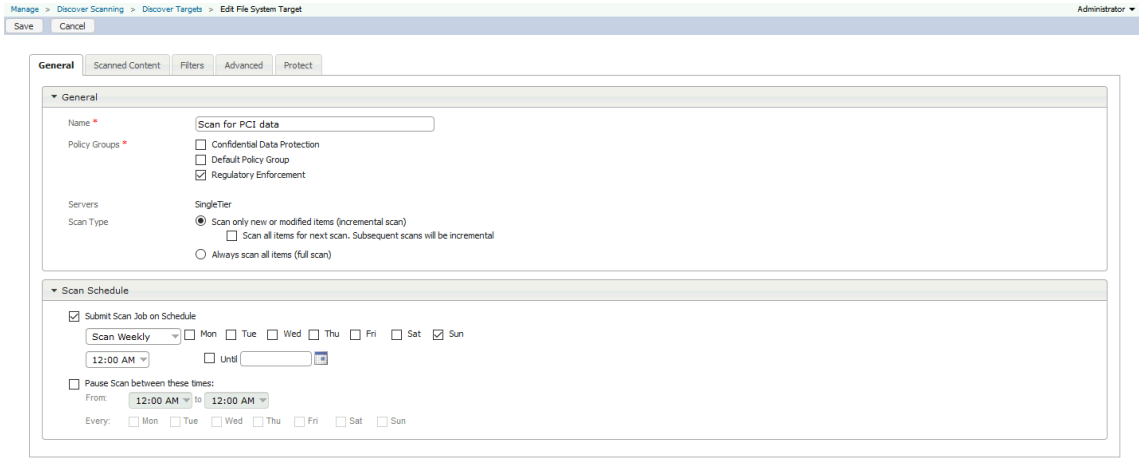

Figure 22 – PCI DSS Discovery Scan

We need to configure the scan with credentials that have access to the share being scanned, as well as specifying the file shares being scanned. No other settings were configured for this scan:

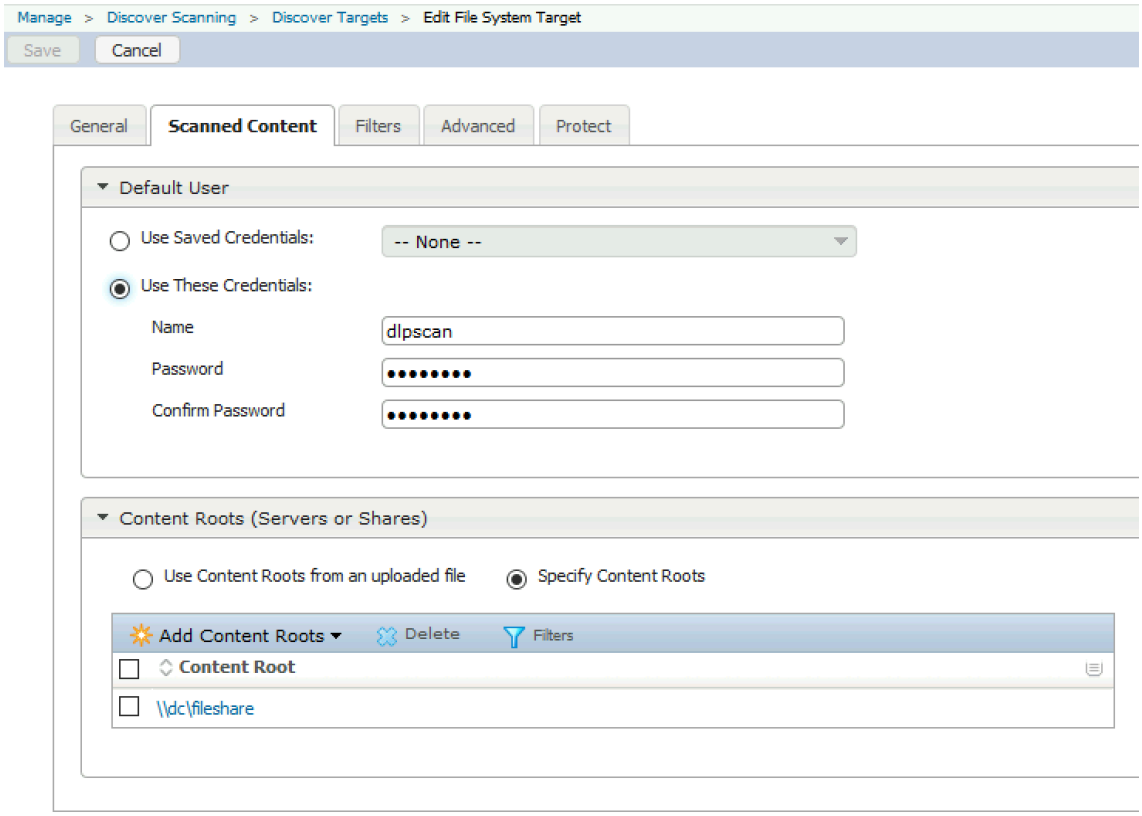

Figure 23 - PCI DSS Discovery Scan Conetnt

Once the scan is started and finished, we can analyze the results:

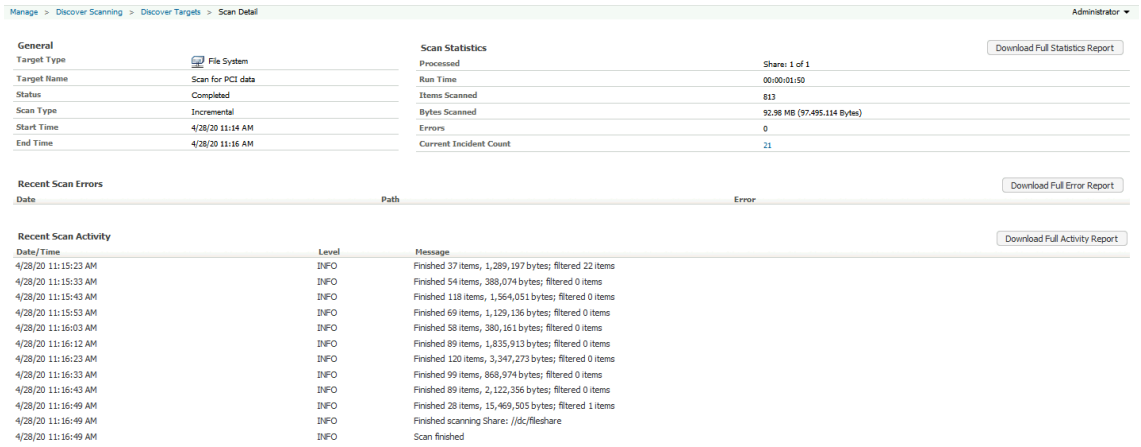

Figure 24 - PCI DSS Discovery Scan Result

This scan identified 21 security incidents which we can see detailed information of what was detected:

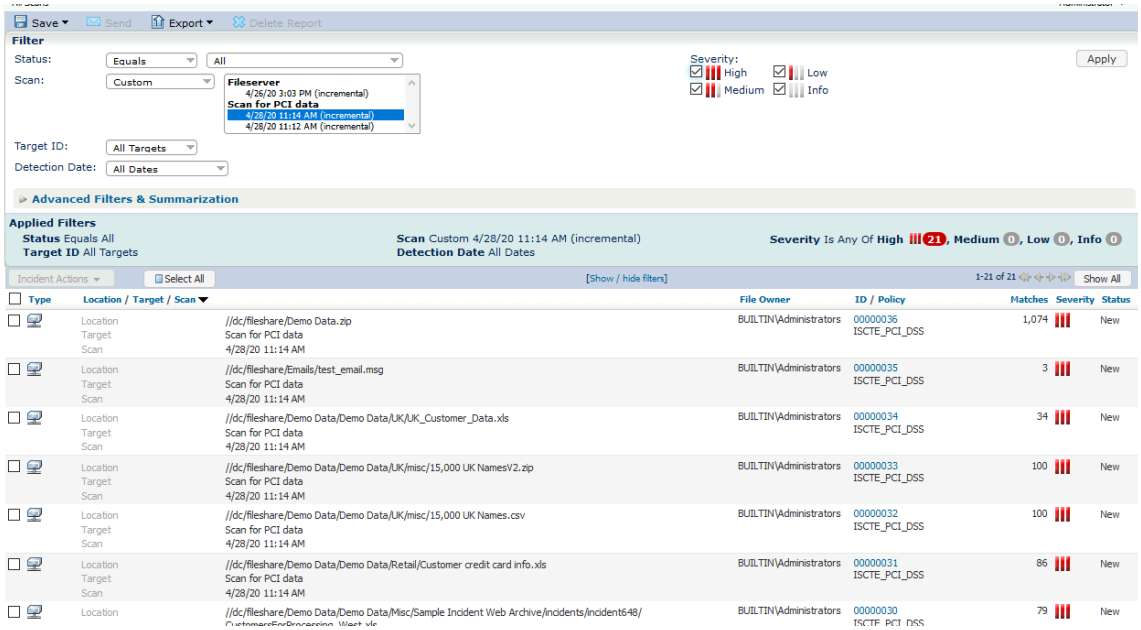

Figure 25 – PCI DSS Discovery Scan Result Details

One of the detected files contains the following:

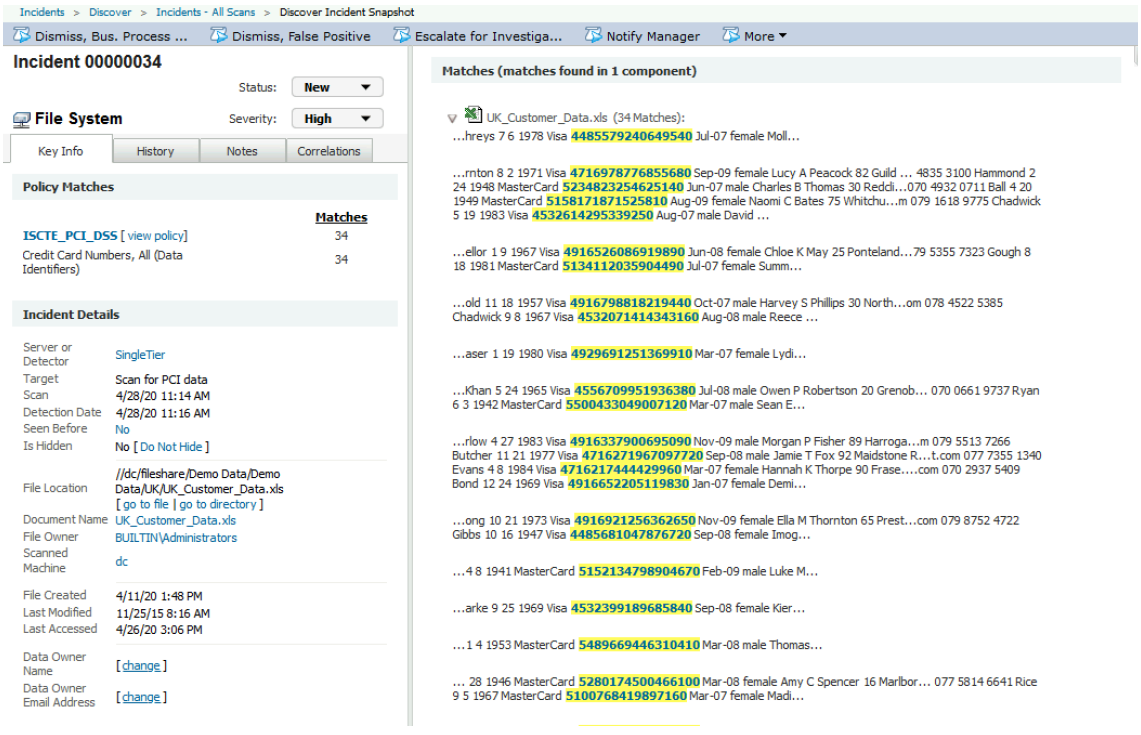

Figure 26 - PCI DSS Discovery Incident Detail

From the detailed incident information, we can clearly identify why this incident was created among other relevant information. By using role-based access control (RBAC) we could hide for specific users some information, this is useful in case I don't want for

the first responder to have access to the content that triggered the policy (matches), but for users that handles escalations, they can see which content triggered the policy.

As a summary for this test case, the focus was identifying confidential information related to PCI DSS stored in network file shares (Data at Rest). By creating a DLP policy we were able to scan the target and identify information at risk. These results provide the evidences that the solution will in fact reduce the risk of data loss. Business unit leaders and data owners can now identify any broken business processes by identifying the incidents generated. Although it was not configured, it is possible to quarantine files or copy them, automatically, thus resulting in an automated remediation.

4.2.2. Test Case 2 – Data Classification Policy

This test case will focus on enforcing the information classification policy. Classifying information is a good practice, because otherwise, all information will have the same value; since some information is more critical than other, by using a classification scheme, specific security controls can be applied to the most critical information. To safeguard confidential information, a process exists, and end-users are required to classify the information as it is described in section 2.5. DLP will ensure the enforcement of the policies. The goal is to prevent that information classified as “Confidential”, “Sensitive” or “Private” cannot leave the organization.

This use case focuses on Data at Rest, Data in Use and Data in Motion and requires the following DLP modules to be implemented:

- DLP Management Console (Enforce), including database
- DLP Endpoint Discover and Prevent
- DLP Network Discover and Protect
- DLP Network for Email

First, a DLP Policy is created to map the classification policy:

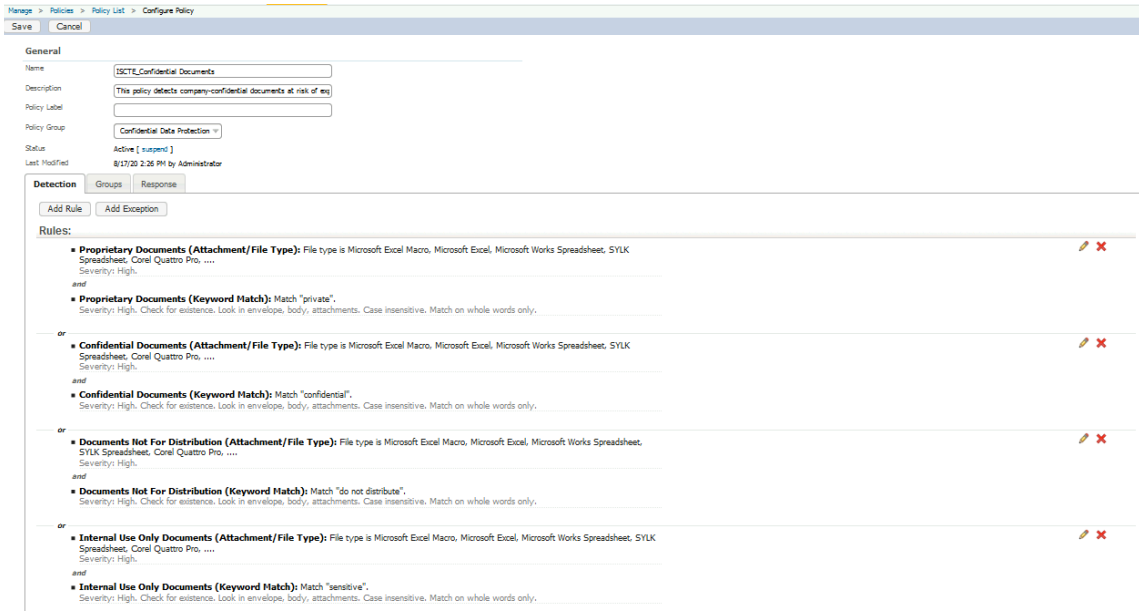

Figure 27 - Data Classification Policy

As the response rules, the following will be executed:

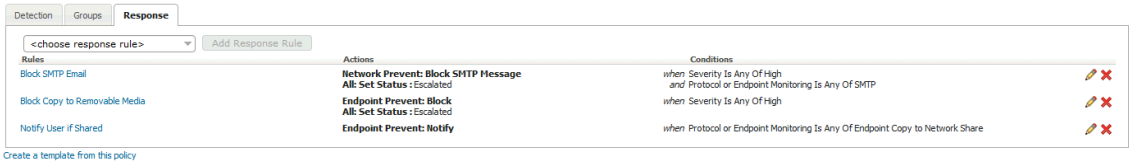

Figure 28 - Data Classification Policy Response

When an email is sent containing confidential information, it will be blocked:

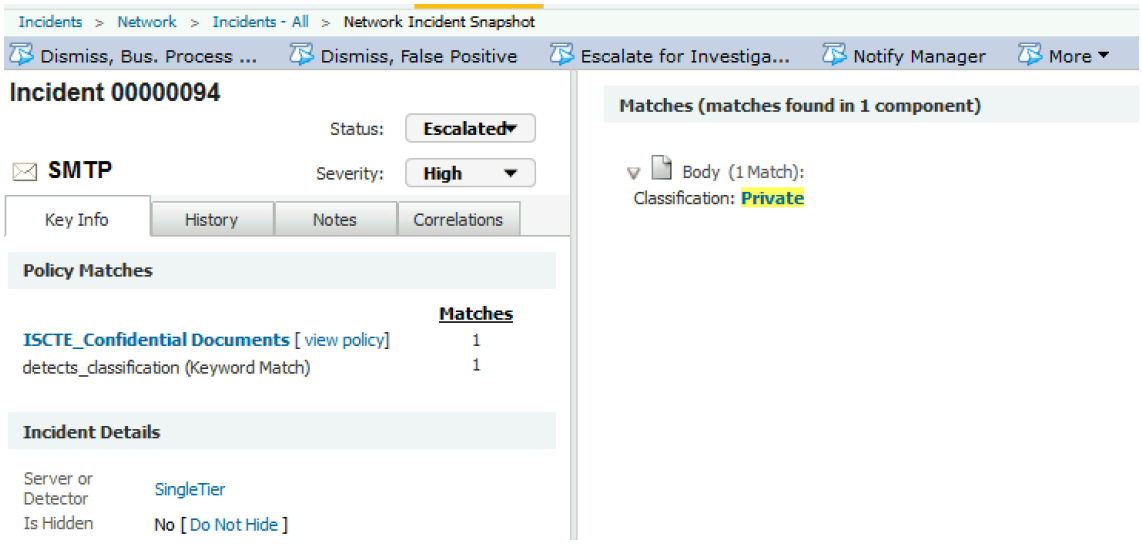

Figure 29 - Data Classification Policy Incident Detail

When a user tries to copy a confidential file from a network share to the endpoint the action is blocked:

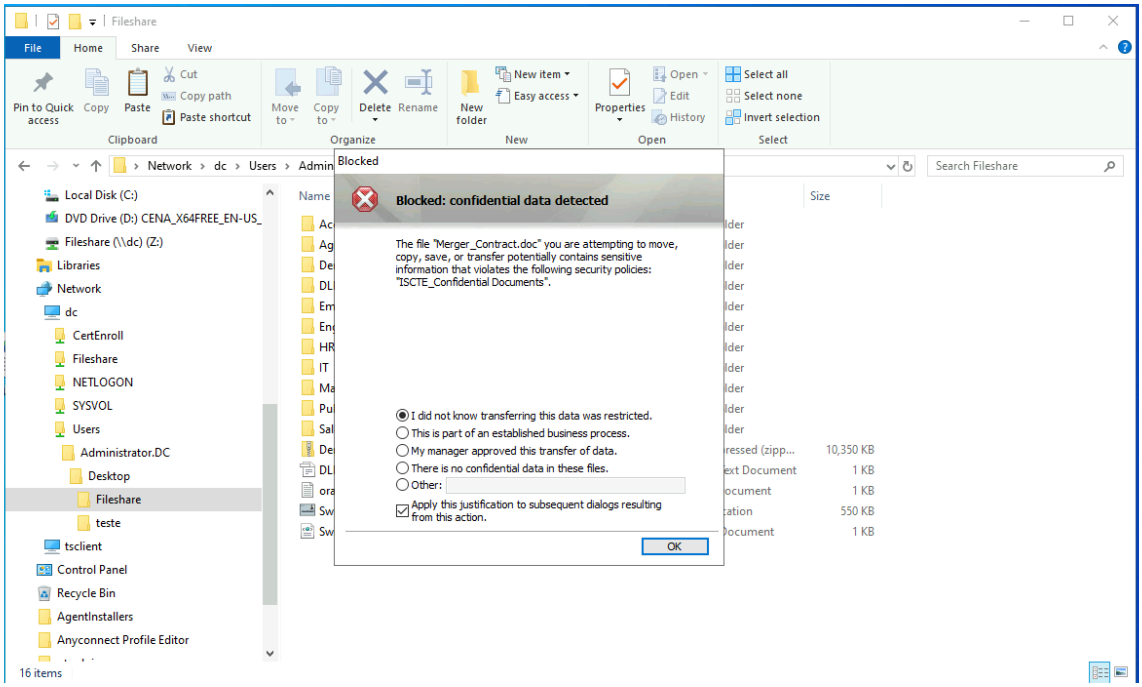

Figure 30 - Data Classification Policy Endpoint Block

Not only the file is blocked from being copied but the user is also notified that this action was blocked. This promotes awareness and reduces risky behaviors.

When performing scans, for example, in a network file share like the example below, it is possible to identify where the confidential files are stored:

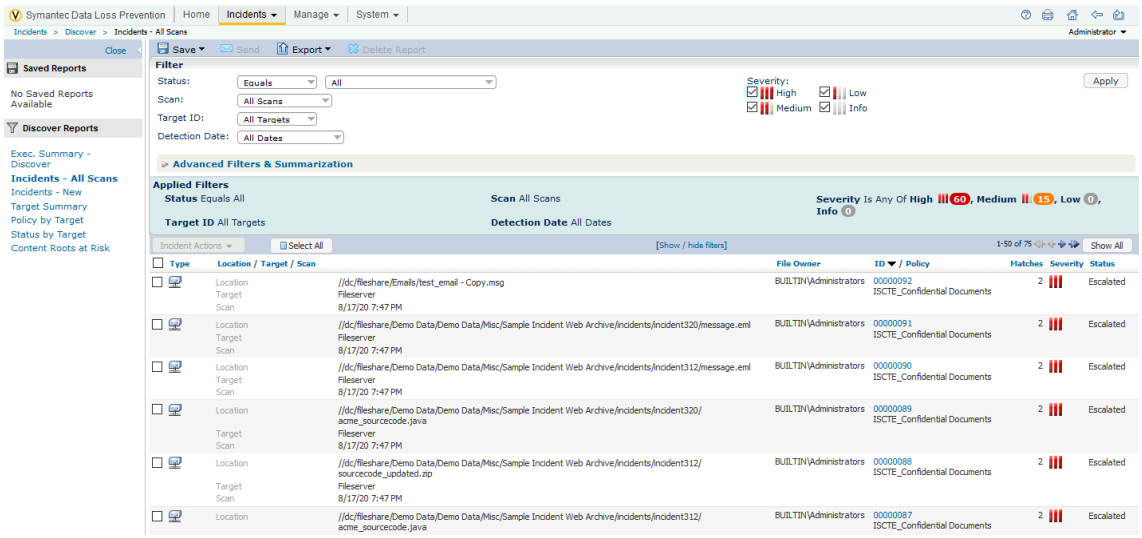

Figure 31 - Data Classification Policy Discovery Scan

An incident responder can then determine if the files should be stored in the location that were found and remediate the incident.

As a summary for this test case, the goal was to enforce the data classification policy and ensure that no confidential information leaves the organization. By configuring a DLP Policy detect specific keywords as described in the data classification policy, it was possible to successfully detect and block data loss.

4.2.3. Test Case 3 – Personal Data

This test case will address the requirement of identifying personal data within the organization. Due to General Data Protection Regulation (GDPR) having visibility on where personal data is stored and used is key in order to better protect it. Different business units may handle personal data and it can be either structured or unstructured data. This test case will focus on protecting the content of structured data.

This test case focuses on Data at Rest, Data in Use and Data in Motion and requires the following DLP modules to be implemented:

- DLP Management Console (Enforce), including database
- DLP Endpoint Discover and Prevent
- DLP Network Discover and Protect

A DLP Policy is created to index a data source (file share), that contains the personal data. Before the policy is created, we need to create a data profile which contains the files to be indexed:

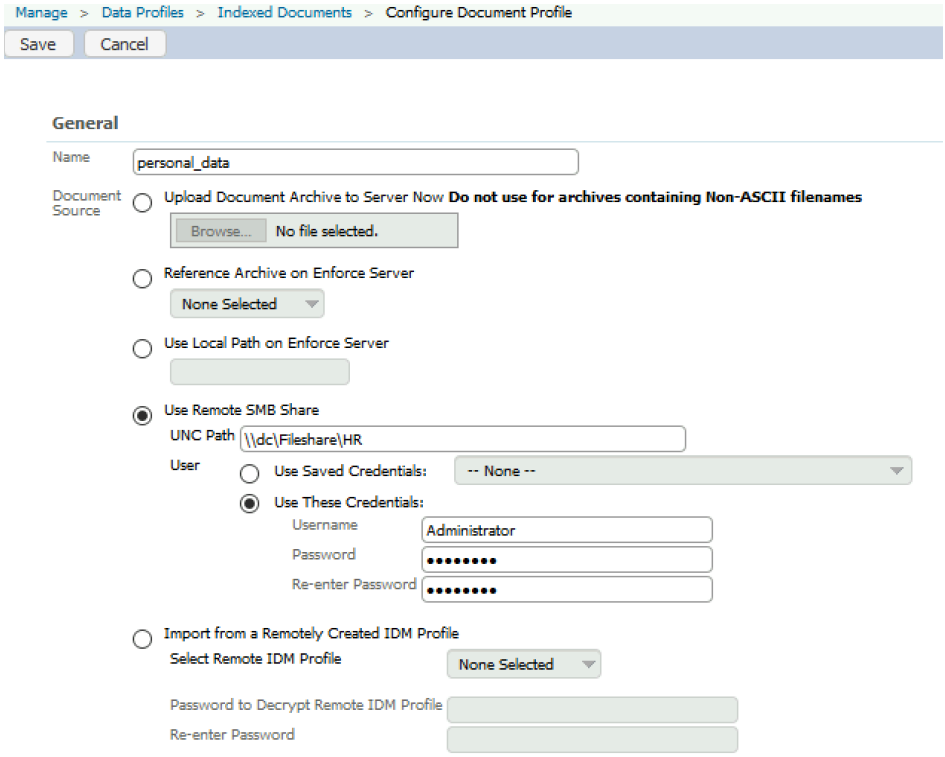

Figure 32 - Personal Data Index

In this scenario, the human resource folder was indexed:

Figure 33 - Personal Data Indexed Files

Next, the DLP Policy is created with the following settings:

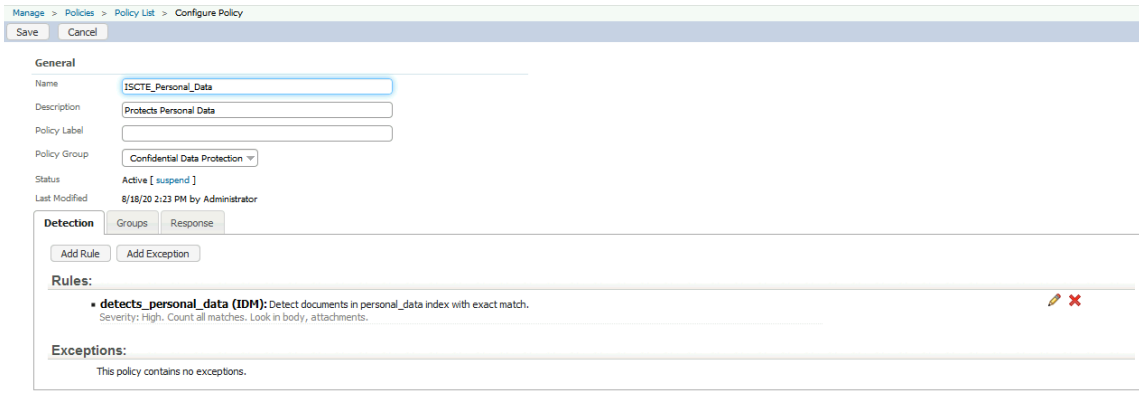

Figure 34 - Personal Data Policy

The response rules configured are as following:

Figure 35 - Personal Data Policy Response

When a user tries to upload to a website, which can be a social media, personal email, cloud storage, etc. the action is blocked:

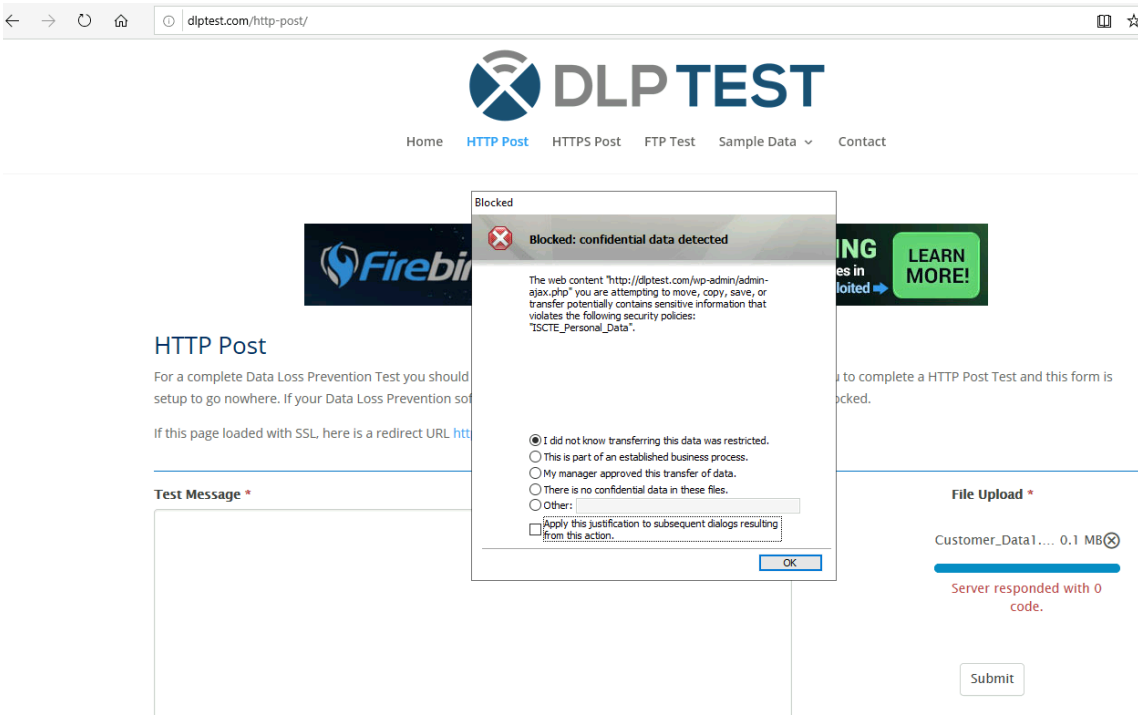

Figure 36 - Personal Data Policy Web HTTP Block

It can either use HTTP like the example above, or HTTPS as the example below:

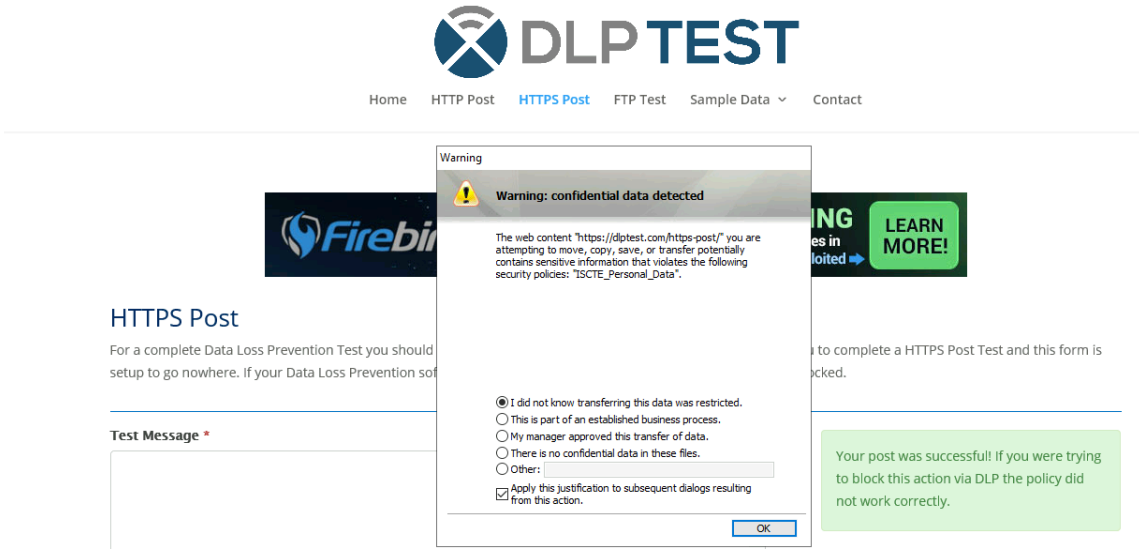

Figure 37 - Personal Data Policy Web HTTP Monitor

In the case of HTTPS, the policy was configured to only notify and not block, however, the action could be blocked as it was HTTPS traffic.

From the DLP Management console, the new incidents are shown:

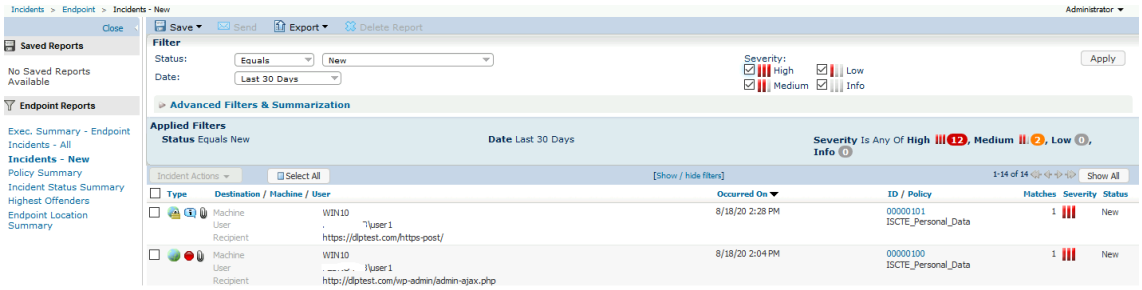

Figure 38 - Personal Data Incidents

When running a scan, personal data will also be identified:

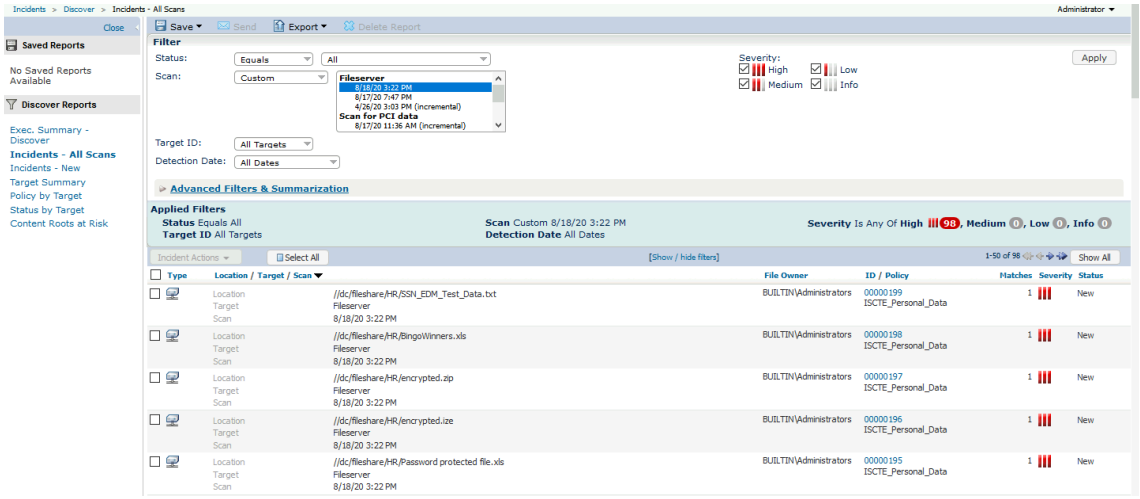

Figure 39 - Personal Data Discovery Scan

As a summary for this test case, the goal was to identify personal information at rest, in use and in transit. By indexing the network share that contained the personal data, a DLP Policy can be created to detect either the whole document of a subset of the document. After a response rule can be applied that will monitor, notify or block the action being taken.

4.2.4. Test Case 4 – Custom Detections

The DLP solution implemented, provides a scripting language to help the development of custom aspects of detection, including file type identification and custom validators that can be used for data identifiers. One reason to use custom detections is to be able to detect specific information not available by default in DLP solutions. Some examples can be Portuguese related information, such as: National ID Number or Social Security Number. An advantage of using the scripting language instead of regular expressions is that we can apply specific validators such as check digits or Luhn algorithm (the latter identifies credit card numbers, among others.).

The objective for this test case is to develop custom detection to detect this kind of information.

First, it will be created a new data identifier in the DLP solution to identify the National ID Number with specific patterns to be identified:

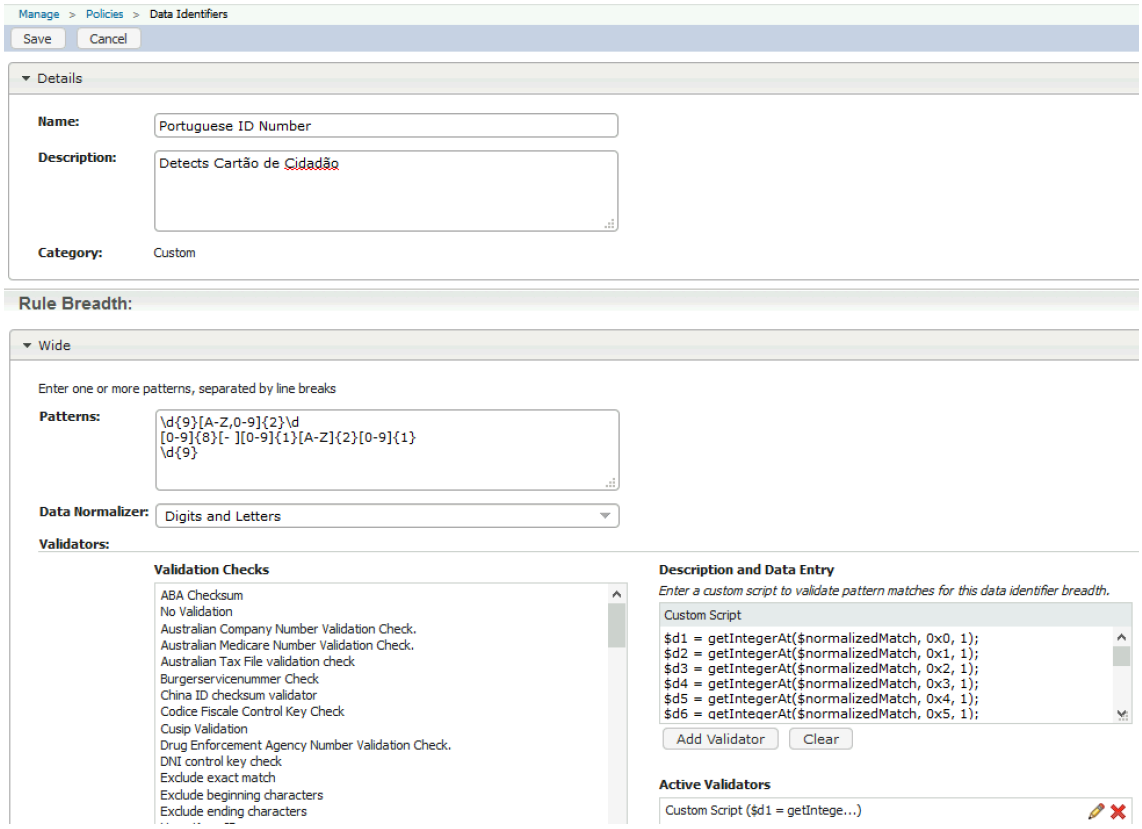

Figure 40 - Portuguese ID Data Identifier

The content of the script is:

```
$d1 = getIntegerAt($normalizedMatch, 0x0, 1);

$d2 = getIntegerAt($normalizedMatch, 0x1, 1);

$d3 = getIntegerAt($normalizedMatch, 0x2, 1);

$d4 = getIntegerAt($normalizedMatch, 0x3, 1);

$d5 = getIntegerAt($normalizedMatch, 0x4, 1);

$d6 = getIntegerAt($normalizedMatch, 0x5, 1);

$d7 = getIntegerAt($normalizedMatch, 0x6, 1);

$d8 = getIntegerAt($normalizedMatch, 0x7, 1);

$c1 = getIntegerAt($normalizedMatch, 0x8, 1);


$iRuleOutAllZeros = add($d1, $d2, $d3, $d4, $d5, $d6, $d7, $d8, $c1);

assertTrue( $iRuleOutAllZeros > 0 );
```

```
$d1 = multiply($d1, 9);

$d2 = multiply($d2, 8);

$d3 = multiply($d3, 7);

$d4 = multiply($d4, 6);

$d5 = multiply($d5, 5);

$d6 = multiply($d6, 4);

$d7 = multiply($d7, 3);

$d8 = multiply($d8, 2);


$iCheckSum = add($d1, $d2, $d3, $d4, $d5, $d6, $d7, $d8);

$iCheckSum = mod($iCheckSum, 11);

$iCheckSum = sub(11, $iCheckSum);


assertTrue($iCheckSum ==$c1);
```

The check digit in this example is being performed in the 9th digit to be able to detect the old citizen card. This can be easily changed by performing the required changes in the Patterns of this data identifier. This script works by:

1. Sum the nine digits and ensure it is bigger than 0
2. Multiply the first 8 digits starting by multiplying the first digit with 9 and ending with the 8th digit with 2
3. Sum the multiplied numbers and calculate remainder
4. The subtraction of the remainder per 11 will give the check digit

This custom validator will be able to detect National ID Numbers in the following format (assuming the check digit is valid): 123761239; 121134524AA1.

Second, we create a new DLP policy to detect this newly created data identifier:

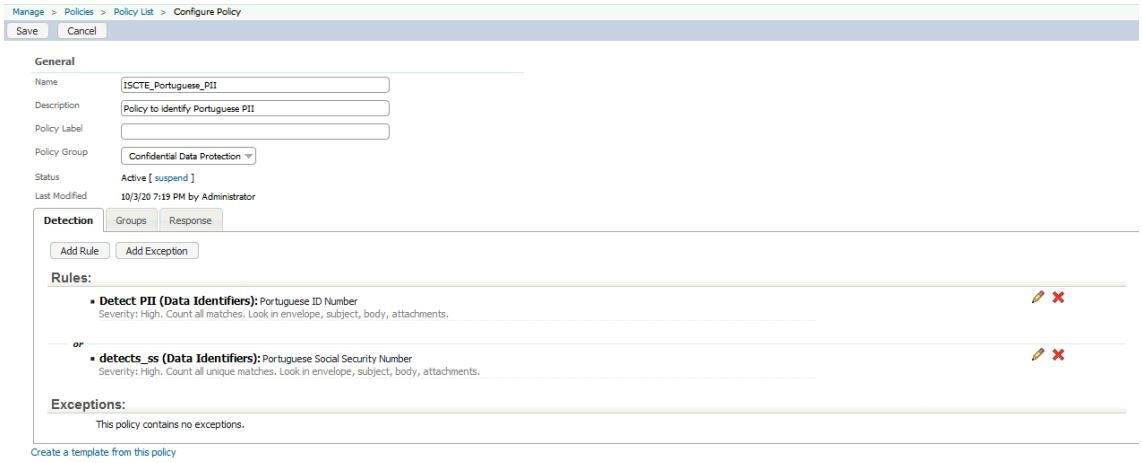

Figure 41 - Portuguese National ID Card Policy

The response rule created for this policy is to notify the end-user when the information is copied from the endpoint to an external location:

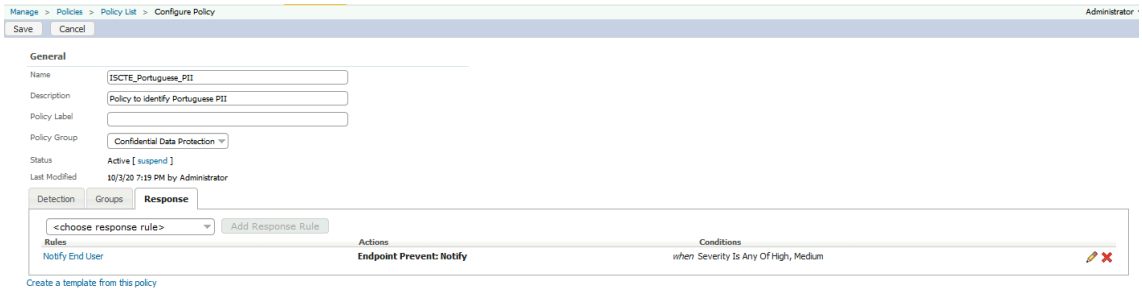

Figure 42 - Portuguese National ID Card Policy, Response Rule

When the user copies information, in this case to a network share, the result is the following:

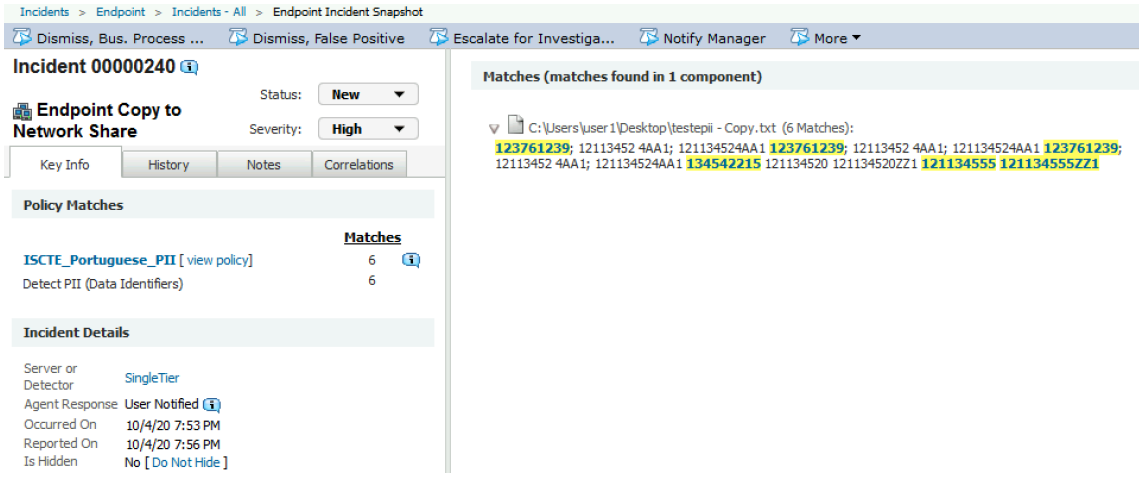

Figure 43 - Portuguese National ID Card Incident

To detect the Portuguese Social Security Number, the process is identical. First, we create a new data identifier:

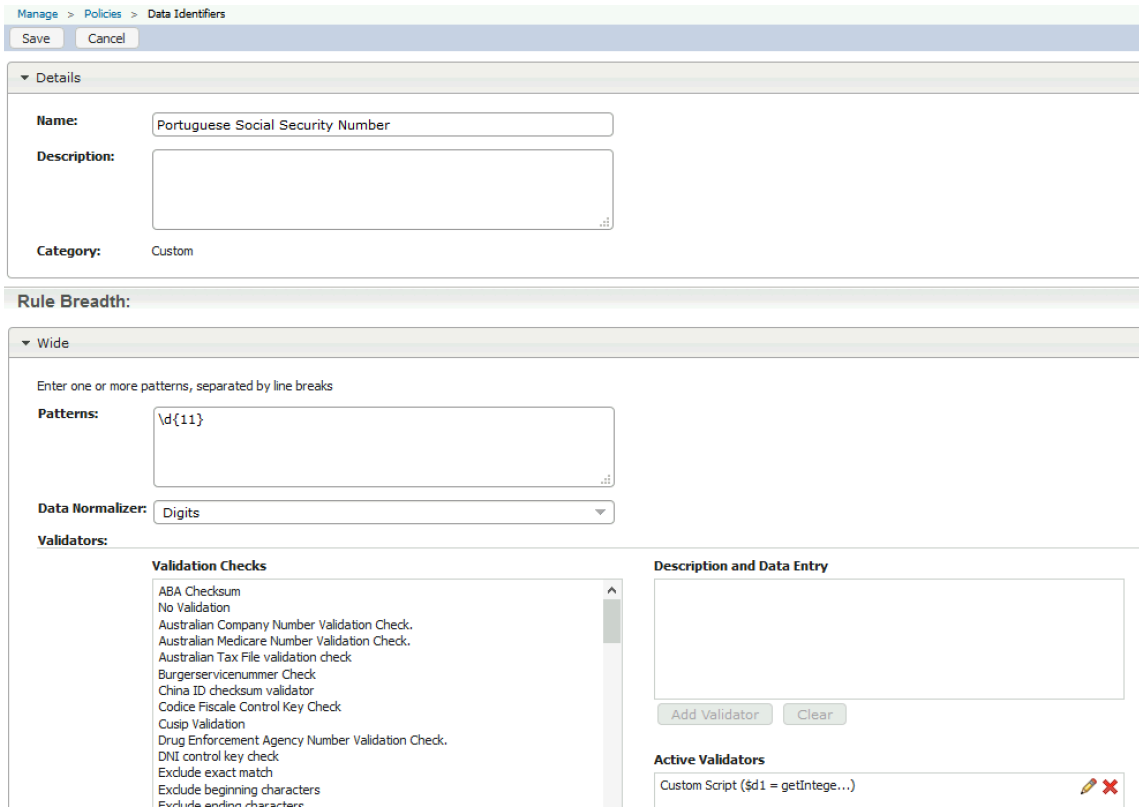

Figure 44 - Portuguese Social Security Number Data Identifier

The content of the script is:

```
$d1 = getIntegerAt($normalizedMatch, 0x0, 1);

$d2 = getIntegerAt($normalizedMatch, 0x1, 1);

$d3 = getIntegerAt($normalizedMatch, 0x2, 1);

$d4 = getIntegerAt($normalizedMatch, 0x3, 1);

$d5 = getIntegerAt($normalizedMatch, 0x4, 1);

$d6 = getIntegerAt($normalizedMatch, 0x5, 1);

$d7 = getIntegerAt($normalizedMatch, 0x6, 1);

$d8 = getIntegerAt($normalizedMatch, 0x7, 1);
```

```
   $d9 = getIntegerAt($normalizedMatch, 0x8, 1);
   $d10 = getIntegerAt($normalizedMatch, 0x9, 1);
   $c1 = getIntegerAt($normalizedMatch, 0xA, 1);

   $iRuleOutAllZeros = add($d1, $d2, $d3, $d4, $d5, $d6, $d7, $d8, $d9,
$d10, $c1);
   assertTrue( $iRuleOutAllZeros > 0 );

   $d1 = multiply($d1, 29);
   $d2 = multiply($d2, 23);
   $d3 = multiply($d3, 19);
   $d4 = multiply($d4, 17);
   $d5 = multiply($d5, 13);
   $d6 = multiply($d6, 11);
   $d7 = multiply($d7, 7);
   $d8 = multiply($d8, 5);
   $d9 = multiply($d9, 3);
   $d10 = multiply($d10, 2);

   $iCheckSum = add($d1, $d2, $d3, $d4, $d5, $d6, $d7, $d8, $d9, $d10);

   $itemp = mod($iCheckSum, 10);
   $iCheckSum = sub(9, $itemp);

   assertTrue($iCheckSum ==$c1);
```

This script works by:

1. Sum the eleven numbers that make the Portuguese Social Security Number and ensure they are not zero.

2. Multiple the first ten numbers with prime numbers (2, 3, 5, 7, 11, 13, 17, 19, 23, 29)
3. Sum the multiplied numbers and calculate remainder
4. The subtraction of the remainder per 9 will give the check digit

This custom validator will be able to detect Social Security Number in the following format (assuming the check digit is valid): 12345678901.

The same policy created for the previous data identifier can be reused:

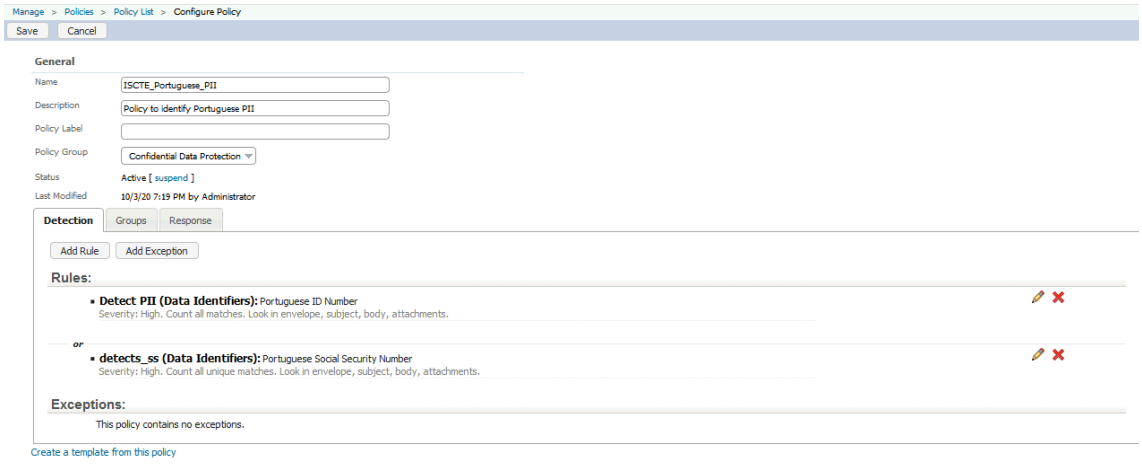

Figure 45 - Portuguese Social Security Number Policy

When the user copies information, in this case to a network share, the result is the following:

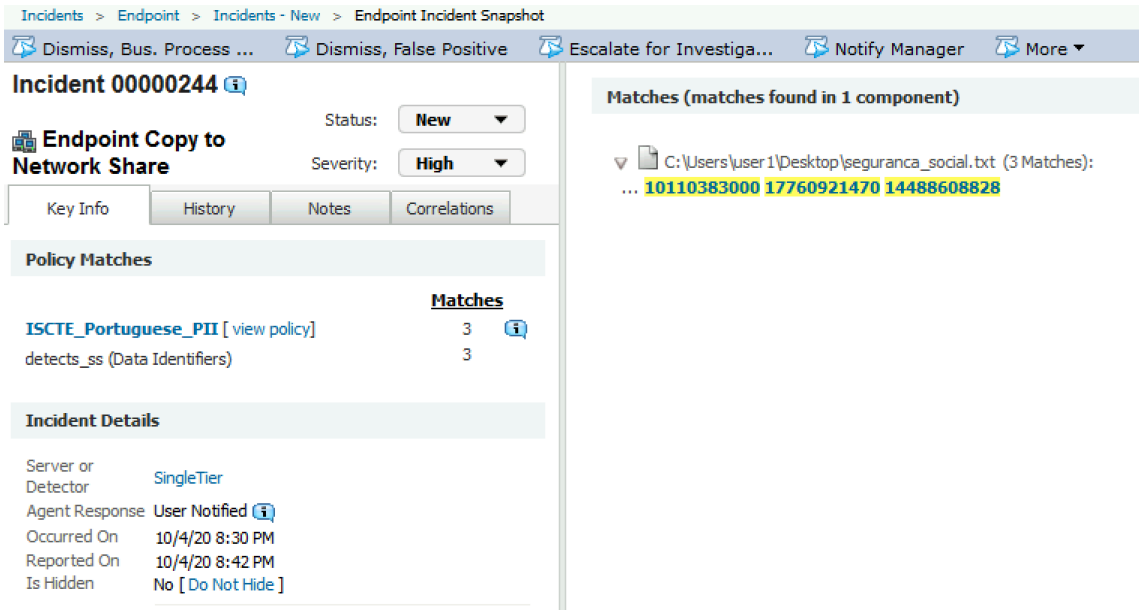

Figure 46 - Portuguese Social Security Number Incident

4.3.Conclusion

This chapter presented a set of test cases to validate whether the proposed architecture reduces the risk of data loss. Various test cases were performed to ensure that data can be protected in different states. Based on the outcome of the test cases it was concluded that the proposed architecture does mitigate the risk of data loss for the proposed scenarios.

Chapter 5 – Conclusion

This chapter concludes this thesis and presents the conclusions of this work that were defined in the first chapter. It also suggests a number of ways for future research to improve knowledge in this area.

5.1.Conclusion

This thesis addressed the topic of data loss prevention, which is an ongoing discussion in the field of information security and validated that it is possible to mitigate and reduce the risk of data loss, by implementing data loss prevention solutions.

Throughout this thesis, both commercial and open-source data loss prevention solutions were investigated to determine how they work, and which are the main components of such solution. Based on this research it was identified that data loss prevention solutions complement existing security controls in place as it focuses on the protection of data itself, therefore it should be part of the overall security strategy. A comparison between multiple solutions allowed to have a better understanding of the key features and integrations.

In order to perform this research, the Design Science Research methodology was used which incorporates a set of phases that results in the creation of an artifact that could be studied and validated. A proposed security architecture was presented in chapter 3 that led to the formulation of the research question. The proposed architecture focuses on the main data states: data in motion, that protects information leaving the organization through the network; data at rest, that protects data stored in file shares or databases and data in use, that is related to the information used in endpoints.

In addition, a commercial DLP solution was implemented and it was highlighted how the key components in the proposed architecture work and how they can be implemented and integrated with the existing infrastructure. One key component is the endpoint agent that gives protection even when the user is outside of the corporate network, thus it was shown in detail how the communication from the endpoint to the server is performed in an encrypted way and by using certificates. To avoid tampering with the endpoint service, a watchdog service is also implemented that continuously monitors the state of the processes. The endpoint agent can also be hidden from the “Windows add/remove

program" section and use a password for both uninstalls and to access DLP endpoint tools.

To validate the proposed security architecture, test cases were defined to show the effectiveness of the solution in preventing data loss. Four test cases were implemented:

1. Protection of PCI DSS data to identify credit card data that might be stored outside of approved locations.
2. Mapping of a data classification policy into a DLP policy. Data classification assigns agreed labels to information based on the level of confidentiality and takes into account the value of the information. If data is classified beforehand it optimizes the value of DLP solutions since it improves its ability to accurately identify data that needs protection.
3. Discovery of personal data across the network.
4. Custom detections, to demonstrate specific customizations that can be implemented to improve the native capabilities of the implemented DLP solution. For this custom detection, it was developed to types of detections: Portuguese National ID and Social Security Numbers.

The test cases have shown that the proposed architecture is able to detect and prevent data loss.

The present research has contributed to the area of DLP by proposing a generic security architecture that can be adapted to different available solutions.

5.2.Future Work

As future work, there are a number of relevant topics that might be addressed. The first one is around cloud computing. Many organizations are moving data into the cloud and adopting cloud files share services, among different workloads. With this transition, data shifts from a centralized model to a decentralized model. Moreover, cloud providers are generally not responsible for customer data, therefore it is important to understand the shared responsibility model of the cloud provider to determine which security tasks are handled by the provider itself and which tasks are handled by the organization. Future research can focus on DLP solutions that provide such native integrations which are able to properly monitor and protect data.

Another topic of interest is to detect and protect information that is already encrypted. Data loss prevention solutions although can apply encryption if sensitive information is identified, are not able to analyze files that are encrypted or use a system for digital rights management, and this poses a risk. Further investigation on methods that allow to decrypt encrypted files, determine whether they are sensitive, apply an action and re-encrypt data is advisable.

Finally, future research on how DLP solutions are used in companies that accepts Bring Your Own Device (BYOD). If an organization allows employees to bring their own computing devices to the workplace, whether it is a smartphone, tablet or a laptop this practice can increase risk. Usually if employees are using personal devices, they might not be able to access sensitive information and if they do, the devices can run endpoint DLP software. Regarding mobile devices in particular and with the fact that some information can be stored in cloud services it translates in increased risk as users can access and download sensitive information to personal mobile devices.

Appendix A: Installing Symantec Data Loss Prevention

In order to install the Symantec DLP Solution, for demonstration purposes, the database (Oracle), Enforce Server (Management Console) and Detection Servers will be installed in the same server (Single Tier Install). It is assumed that the database is already installed and available.

To perform a successfully installation the DLP Enforce Server or a Detection server:

Step 1: Click next of the welcome window.

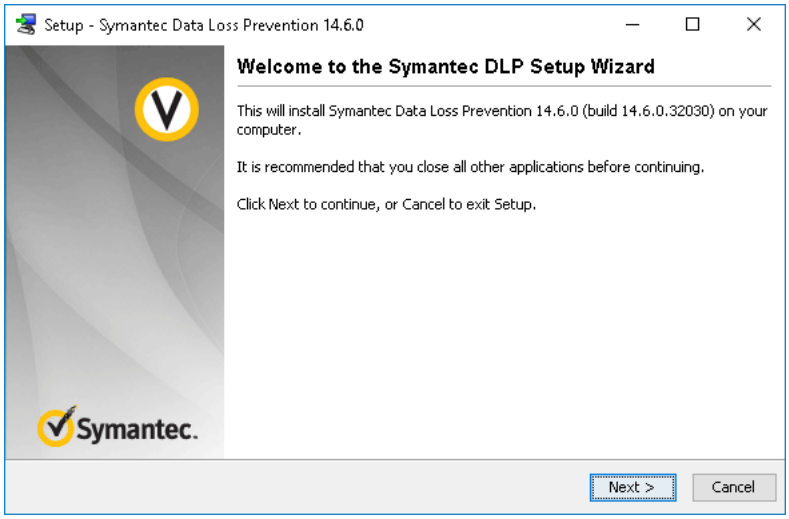

Figure 47 - DLP Install, Step 1

Step 2: Accept the license agreement and click next.

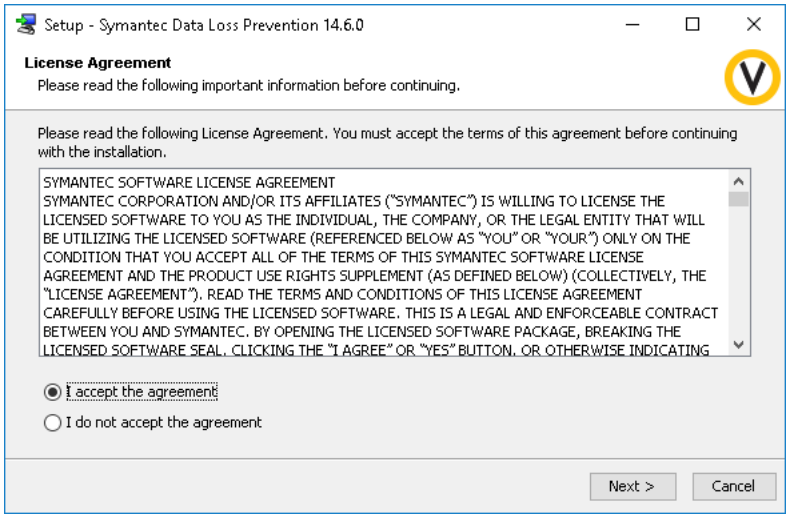

Figure 48 - DLP Install, Step 2

Step 3: Select the components you wish to install, in this case a single tier which means that a single server will host all the components.

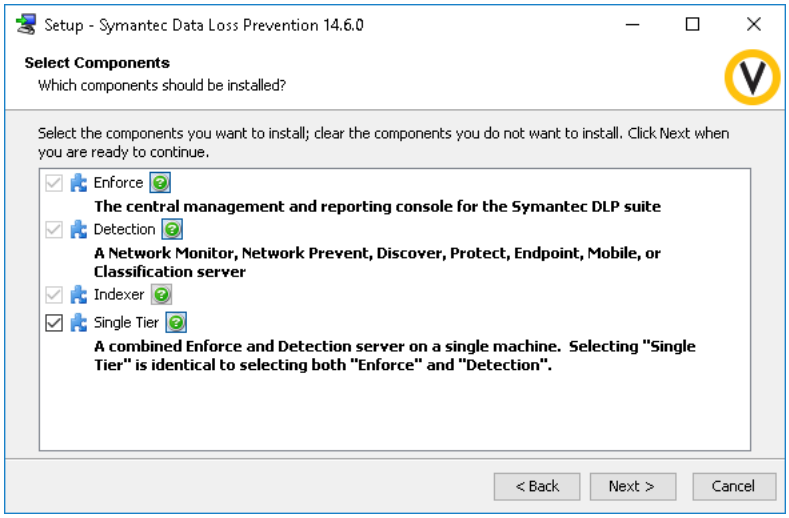

Figure 49 - DLP Install, Step 3

Step 4: Select the license file and click next.

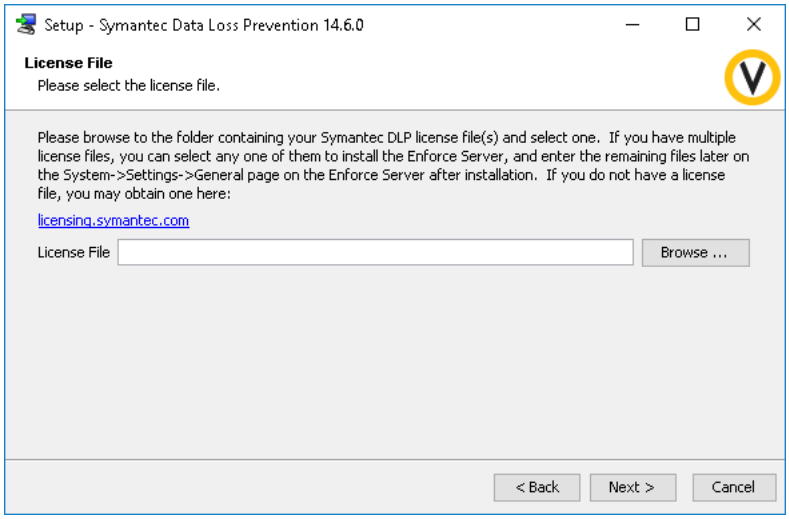

Figure 50 - DLP Install, Step 4

Step 5: Select next in the WinPcap windows. If this server will have the role of Network Monitor is recommended to install WinPcap.

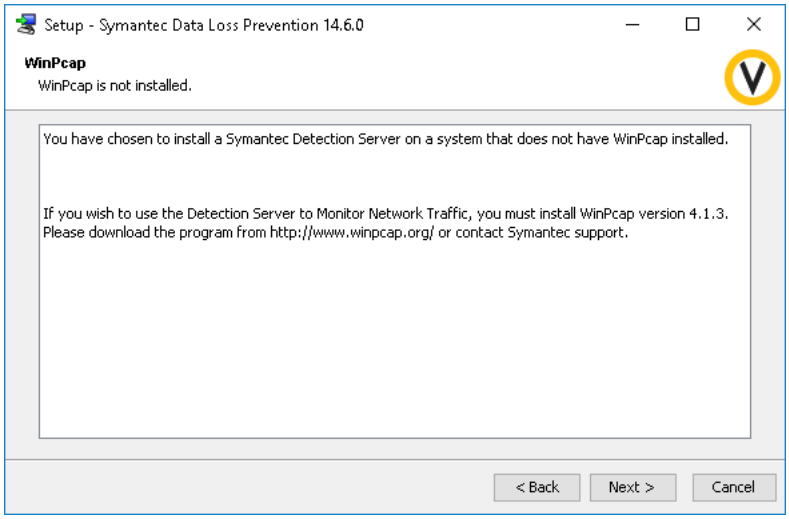

Figure 51 - DLP Install, Step 5

Step 6: Select the installation directory and click next.

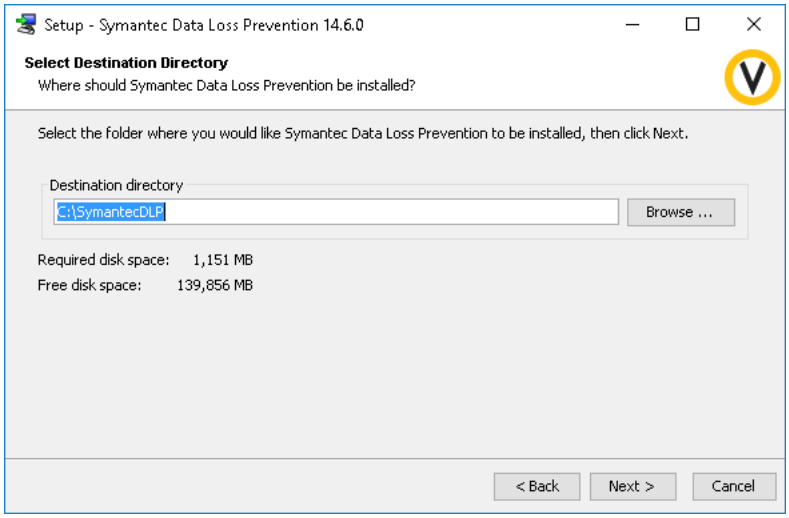

Figure 52 - DLP Install, Step 6

Step 7: Click next on the select start menu folder window.

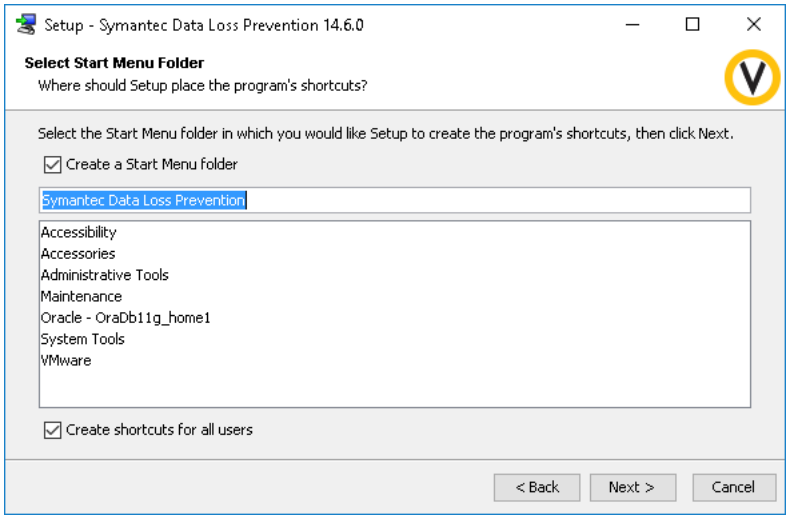

Figure 53 - DLP Install, Step 7

Step 8: If you don't have a service account created, accept the default option and click next.

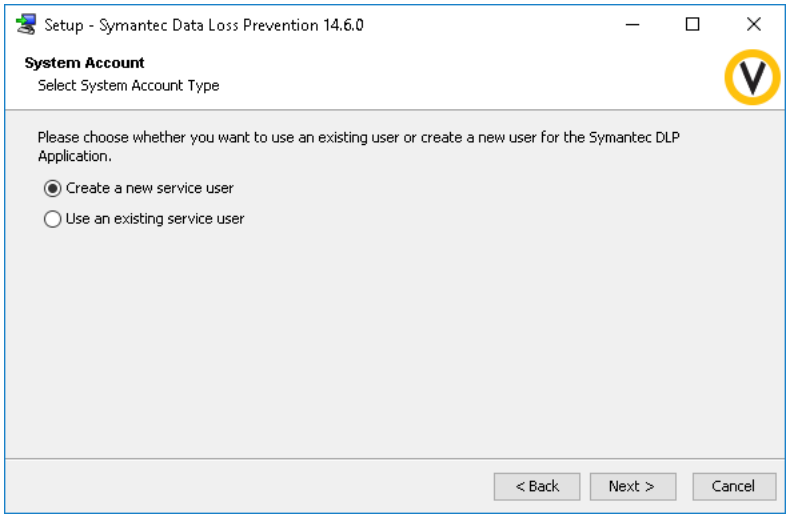

Figure 54 - DLP Install, Step 8

Step 9: Select a username and password and click next.

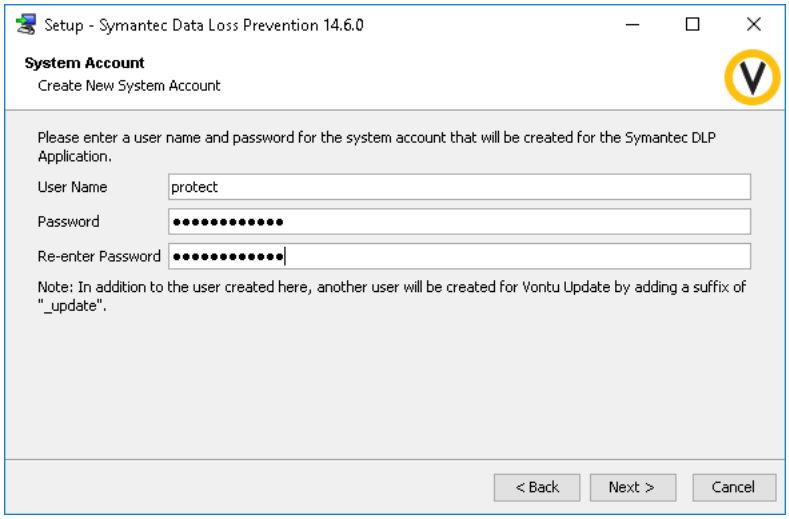

Figure 55 - DLP Install, Step 9

Step 10: Accept the default transport configuration and click next.

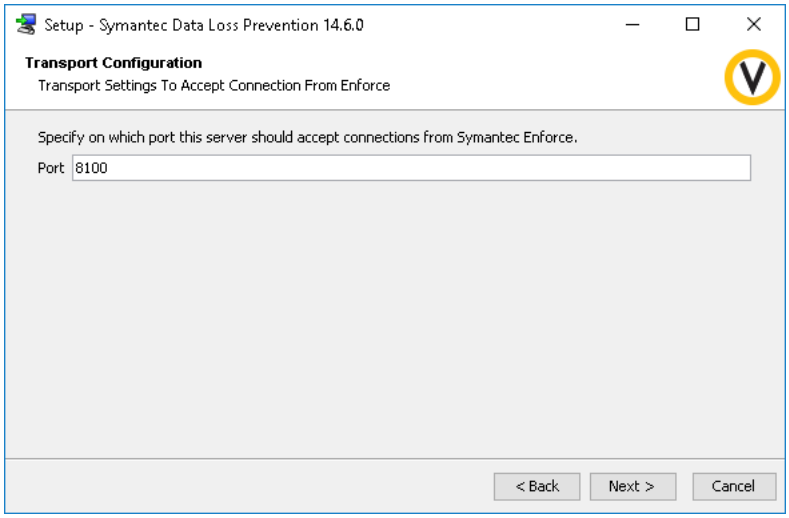

Figure 56 - DLP Install, Step 10

Step 11: Enter the IP Address of the Oracle Database Server and Listener Port and click next.

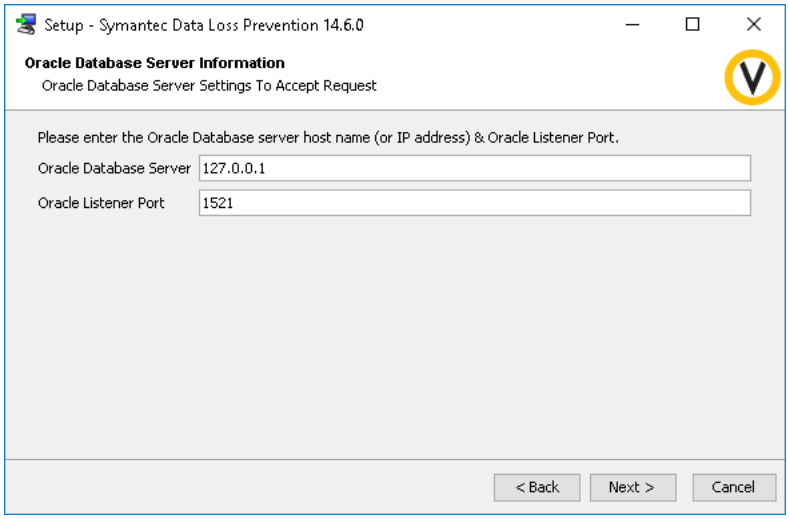

Figure 57 - DLP Install, Step 11

Step 12: Enter the username, password and oracle SID and click next.

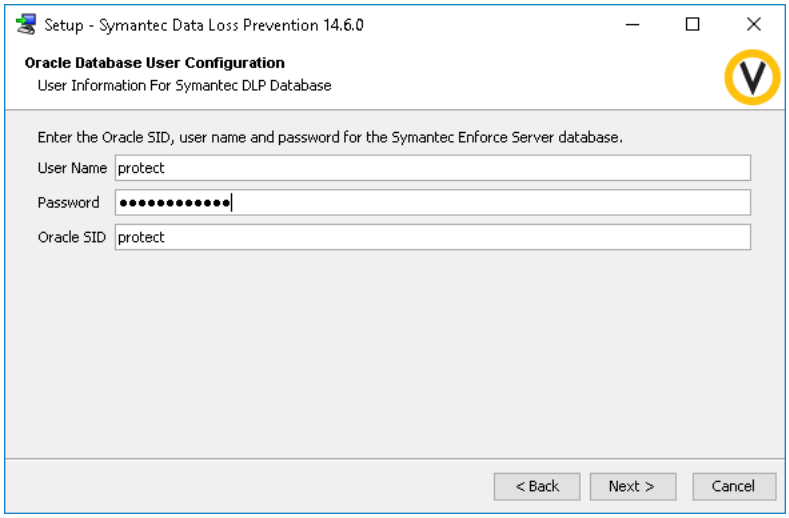

Figure 58 - DLP Install, Step 12

Step 13: Click next on the additional locale window.

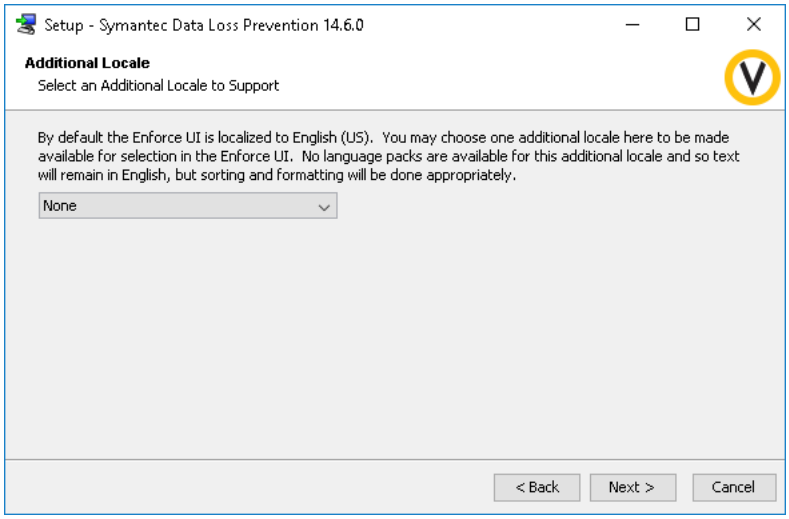

Figure 59 - DLP Install, Step 13

Step 14: Click next on the initialize DLP Database.

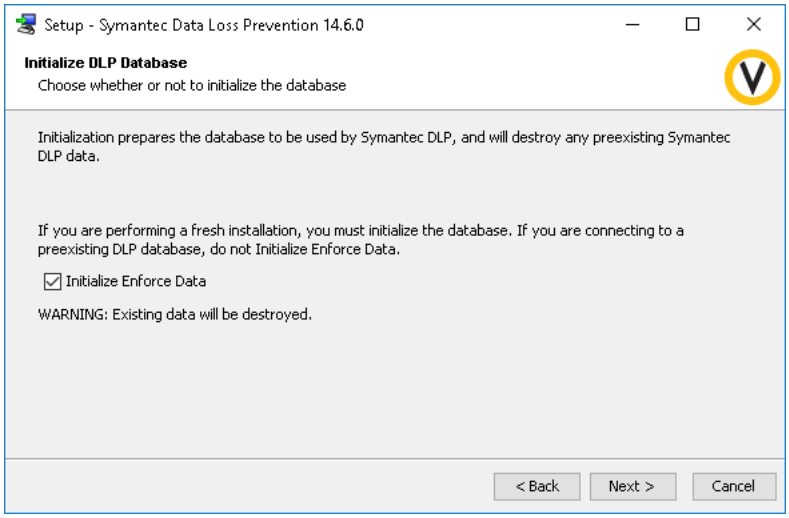

Figure 60 - DLP Install, Step 14

Step 15: Accept the default password authentication only and click next.

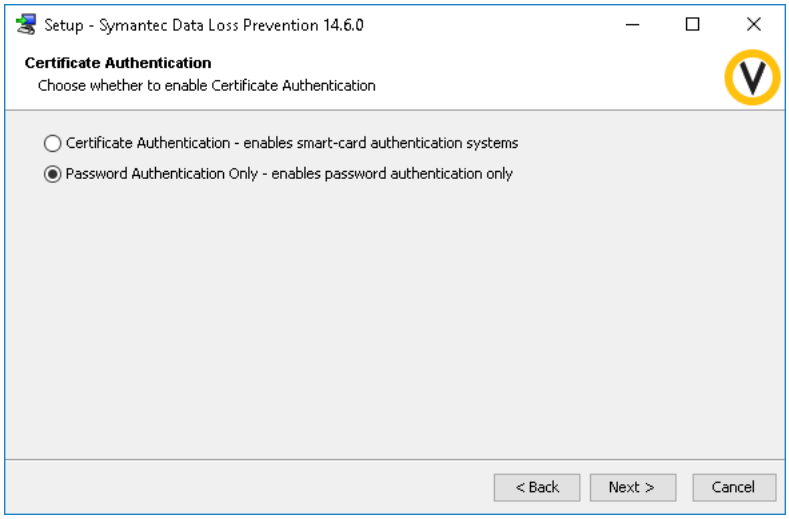

Figure 61 - DLP Install, Step 15

Step 16: Enter the Administrator user password and click next.

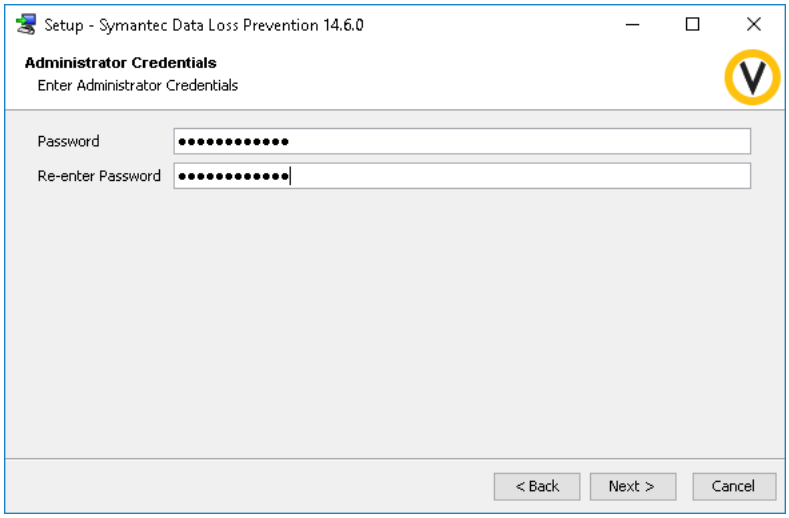

Figure 62 - DLP Install, Step 16

Step 17: Click next on the enable external storage window.

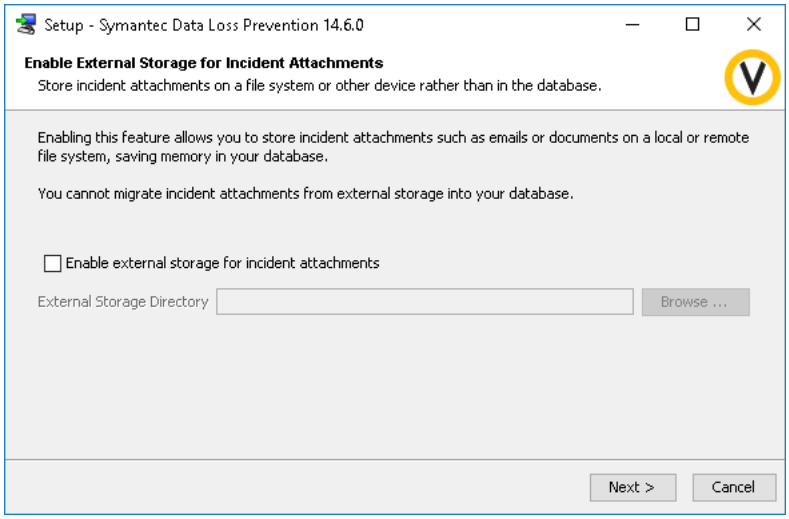

Figure 63 - DLP Install, Step 17

Step 18: Select participate in supportability telemetry if you wish and select whether this is a production or test system.

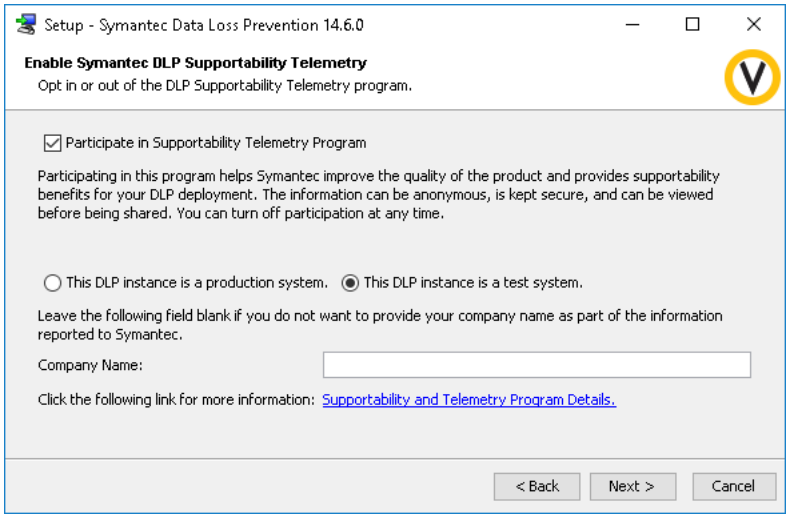

Figure 64 - DLP Install, Step 18

Step 19: Click finish of the completing the Symantec Data Loss Prevention Setup Wizard window.

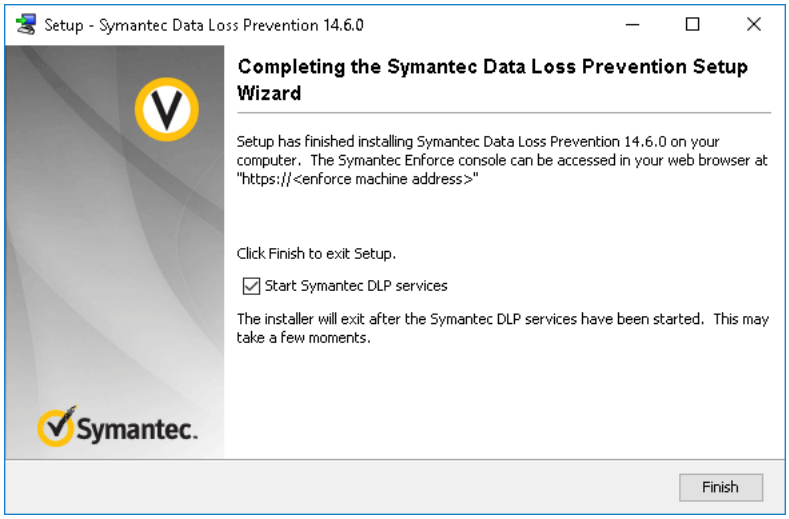

Figure 65 - DLP Install, Step 19

Step 20: After the installation it is recommended to import a solution pack. To do so, stop the VontuManager in the bin directory on which Symantec DLP Enforce Server was installed.

Figure 66 - DLP Install, Step 20

Step 21: Run the import command as shown below.

Figure 67 - DLP Install, Step 21

Step 22: Start the VontuManager service.

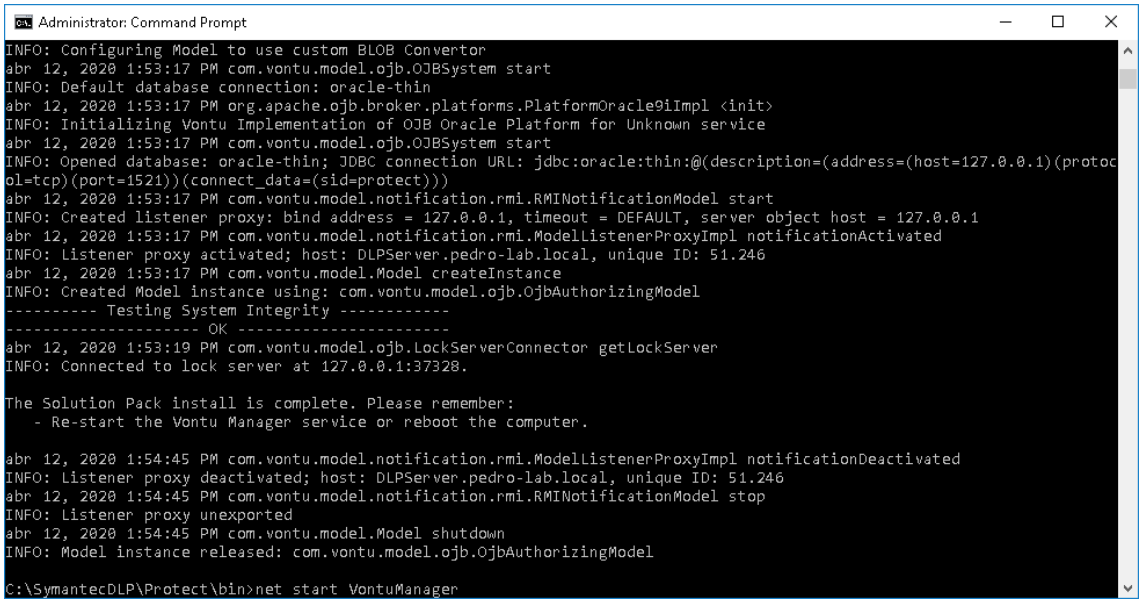

Figure 68 - DLP Install, Step 22

Step 23: Open a browser window to the fully qualified domain name of the Enforce Server, enter the Administrator username and password.

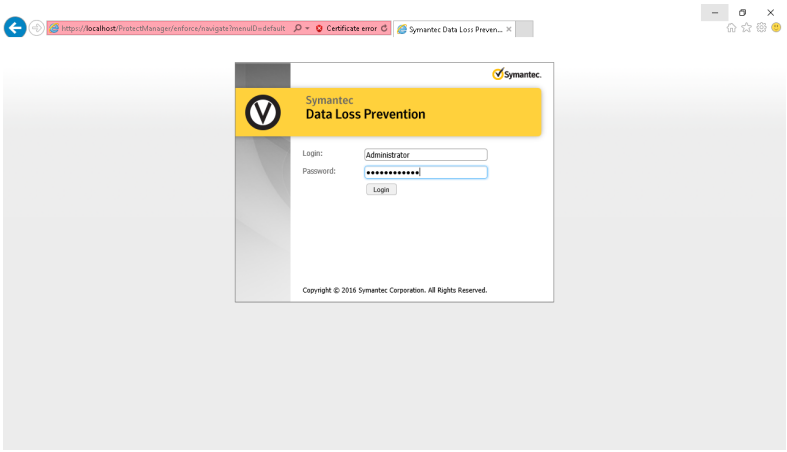

Figure 69 - DLP Install, Step 23

Step 24: Enter a name, title and company and click next.

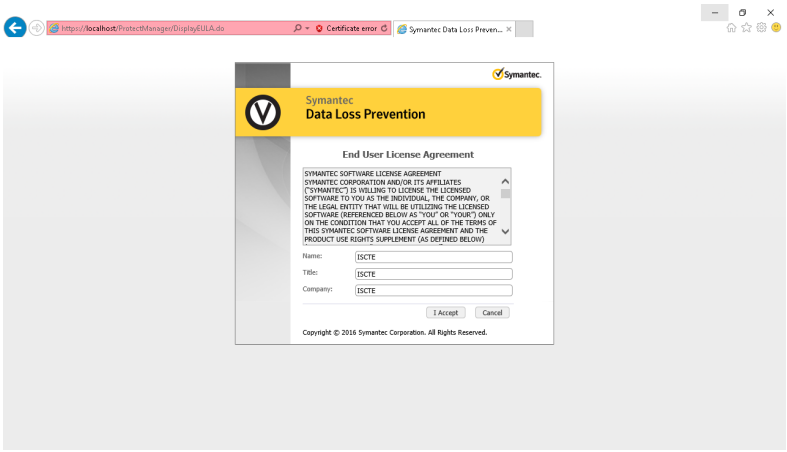

Figure 70 - DLP Install, Step 24

Step 25: The installation is completed, and the home dashboard is presented.

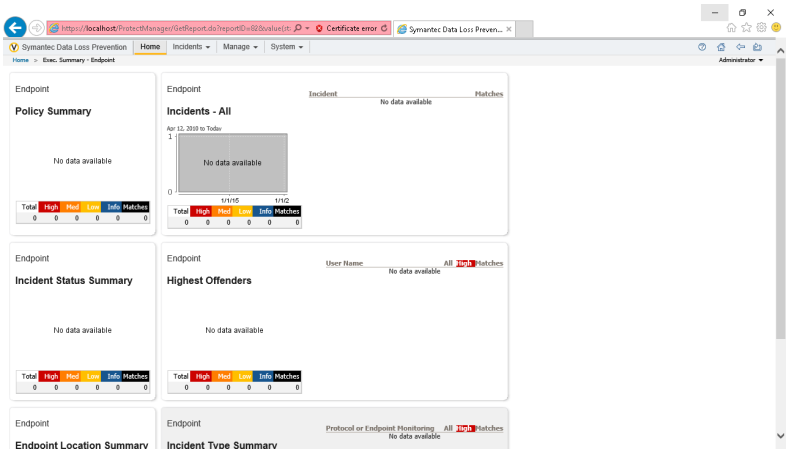

Figure 71 - DLP Install, Step 25

References

Cisco. (2019). What Is DLP? - Data Loss Prevention - Cisco. Retrieved January 5, 2019, from https://www.cisco.com/c/en/us/products/security/email-security-appliance/data-loss-prevention-dlp.html

Ouellet. (2009). Magic Quadrant for Content-Aware Data Loss Prevention. Retrieved from https://www.symantec.com/content/en/us/about/media/industryanalysts/Gartner_DLP_2009_MQ.pdf

Reed, B., & Wynne, N. (2017). Magic Quadrant for Data Loss Prevention. Retrieved from https://www.gartner.com/doc/reprints?id=1-3TOGM5S&ct=170216&st=sb&mkt_tok=eyJpIjoiTlRRMU9UTmlNakJqWW1ZdyIsInQiOiJQSHF4V3BjNjBYU3dha2hxdWxPWkRnRG5QaU0yMVFNMDZaY0VvODVjcmdlNFwvKzhYS3RqYTJSYis2czJ1Z21qemJBNDlqWHBwTGNQbjljREYwYm9EXC8xNUVPWlJQWThhMGZlVWtqMFBDb

Radicati. (2017). Data Loss Prevention - Market Quadrant 2017. Retrieved from https://www.tecnozero.com/wp-content/uploads/2018/06/informe-radicati-data-loss-prevention-2017.pdf

Reinsel, D., Gantz, J., & Rydning, J. (2018). The Digitization of the World From Edge to Core.

Andress, J., & Winterfeld, S. (2014). The Basics of Information Security: Understanding the Fundamentals of InfoSec in Theory and Practice: Second Edition. The Basics of Information Security: Understanding the Fundamentals of InfoSec in Theory and Practice: Second Edition.

O'Hanley, R., & Tiller, J. (2013). Information Security Management Handbook. In Information Security Management Handbook, Sixth Edition, Volume 7. https://doi.org/10.1201/b15440-45

Baby, A. & Krishnan, H. (2017). A Literature Survey on Data Leak Detection And Prevention Methods. International Journal of Advanced Research in Computer Science, 8(5), 2416-2418.

Shabtai, A., Elovici, Y., & Rokach, L. (2012). A survey of data leakage detection and prevention solutions. Springer.

Verizon. (2018). 2018 Data Breach Investigations Report - 11th edition. Verizon Business Journal.

T. Wüchner and A. Pretschner, "Data Loss Prevention Based on Data-Driven Usage Control," 2012 IEEE 23rd International Symposium on Software Reliability Engineering, Dallas, TX, 2012, pp. 151-160, doi: 10.1109/ISSRE.2012.10.

S. Liu and R. Kuhn, "Data Loss Prevention," in IT Professional, vol. 12, no. 2, pp. 10-13, March-April 2010, doi: 10.1109/MITP.2010.52.

Harris, S. (2010). CISSP All-in-One Exam Guide, Fifth Edition

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144. https://doi.org/10.1016/J.IJINFOMGT.2014.10.007

Stewart, J. M., Chapple, M., & Gibson, D. (2015). CISSP : certified information systems security professional study guide.

Rocha, A., Correia, A. M., Costanzo, S., & Reis, L. P. (Eds.). (2015). New Contributions in Information Systems and Technologies (Vol. 353). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-16486-1

Rich Mogull. (2014). Inside DLP: Full-suite products, DLP lite, content analysis. Retrieved January 9, 2019, from https://searchsecurity.techtarget.com/tip/Inside-DLP-Full-suite-products-DLP-lite-content-analysis

Alzhrani, K., Rudd, E. M., Boult, T. E., & Chow, C. E. (2016). Automated big text security classification. In IEEE International Conference on Intelligence and Security Informatics: Cybersecurity and Big Data, ISI 2016. https://doi.org/10.1109/ISI.2016.7745451

Hart, M., Manadhata, P. K., Johnson, R., & Manadhata, P. (2011). Text Classification for Data Loss Prevention. Retrieved from http://www.hpl.hp.com/techreports/2011/HPL-2011-114.pdf

Goyvaerts, J., & Levithan, S. (2012). Regular expressions cookbook. O'Reilly Media.

Costante, E., Fauri, D., Etalle, S., Hartog, J. Den, & Zannone, N. (2016). A Hybrid Framework for Data Loss Prevention and Detection. In 2016 IEEE Security and Privacy Workshops (SPW) (pp. 324–333). IEEE. https://doi.org/10.1109/SPW.2016.24

Costante, E., Hartog, J., Petkovic, M., Etalle, S., & Pechenizkiy, M. (2016). A white-box anomaly-based framework for database leakage detection. Journal of information security and applications. doi: 10.1016/j.jisa.2016.10.001.

Oriyano, S.-P. (2016). CEH v9: Certified ethical hacker version 9, study guide.

Ponemon Institute. (2018). 2018 Cost of a Data Breach Study: Global Overview. Retrieved from https://www.ibm.com/account/reg/us-en/signup?formid=urx-33316

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). DESIGN SCIENCE IN INFORMATION SYSTEMS RESEARCH 1. Design Science in IS Research MIS Quarterly (Vol. 28). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.1725&rep=rep1&type=pdf

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of Management Information Systems. https://doi.org/10.2753/MIS0742-1222240302

Shapira, Y., Shapira, B., & Shabtai, A. (2013). Content-based data leakage detection using extended fingerprinting. Retrieved from https://www.researchgate.net/publication/235427338_Content-based_data_leakage_detection_using_extended_fingerprinting

Alneyadi, S., Sithirasenan, E., & Muthukkumarasamy, V. (2015). Detecting Data Semantic: A Data Leakage Prevention Approach. In 2015 IEEE Trustcom/BigDataSE/ISPA (pp. 910–917). IEEE. https://doi.org/10.1109/Trustcom.2015.464

Kaur, Kamaljeet & Gupta, Ishu & Singh, Ashutosh. (2017). A Comparative Evaluation of Data Leakage/Loss Prevention Systems (DLPS). 87-95. 10.5121/csit.2017.71008.

Noble, A. P., Kopaee, R., Melek, A., & Nirvik, N. (2010). Data Leak Prevention. ISACA, (September), 1–14. Retrieved from http://www.isaca.org/Knowledge-Center/Research/Documents/DLP-WP-14Sept2010-Research.pdf

Katz, G., Elovici, Y., & Shapira, B. (2014). CoBAn: A context based model for data leakage prevention. Information Sciences, 262, 137–158. https://doi.org/10.1016/j.ins.2013.10.005

Ponemon Institute. (2017). 2017 Cost of Data Breach Study, Global Overview. IBM Security.
https://www.ibm.com/downloads/cas/ZYKLN2E3#:~:text=IBM%20Security%20and%20Ponemon%20Institute,from%20%244.00%20to%20%243.62%20million2

IBM: Cost of a Data Breach Report 2019. (2019). Computer Fraud & Security. https://doi.org/10.1016/s1361-3723(19)30081-8

Garrison, C. & Ncube, M. (2011). A longitudinal analysis of data breaches. Information Management & Computer Security 19(4), 216-230.

Layton, R. & Watters, P. (2014). A methodology for estimating the tangible cost of data breache. Applications, 19(6), 321-330.

Thomas, K., Li, F., Zand, A., Barrett, J., Ranieri, J., Invernizzi, L., ... Bursztein, E. (2017). Data Breaches, Phishing, or Malware? Understanding the Risks of Stolen Credentials. CCS'17, October 30-November 3, 2017, Dallas, TX, USA.

Tahboub, R., & Saleh, Y. (2014). Data leakage/loss prevention systems (DLP). In 2014 World Congress on Computer Applications and Information Systems, WCCAIS 2014. https://doi.org/10.1109/WCCAIS.2014.6916624

Scarfone, K., & Mell, P. (2007). Guide to Intrusion Detection and Prevention Systems (IDPS). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-94

Ashoor, A. S., & Gore, S. (2011). Difference between Intrusion Detection System (IDS) and Intrusion Prevention System (IPS). In Communications in Computer and Information Science. https://doi.org/10.1007/978-3-642-22540-6_48

Liao, H. J., Richard Lin, C. H., Lin, Y. C., & Tung, K. Y. (2013). Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications. https://doi.org/10.1016/j.jnca.2012.09.004

Moraes, A. (2011). Cisco firewalls. Cisco Firewalls.

Cisco. (2017). What Is a Firewall? Cisco. Cisco. Retrieved from https://www.cisco.com/c/en/us/products/security/firewalls/what-is-a-

firewall.html%0Ahttp://www.cisco.com/c/en/us/products/security/firewalls/what-is-a-firewall.html

Koret, J., & Bachaalany, E. (2015). The Antivirus Hacker's Handbook. The Antivirus Hacker's Handbook. https://doi.org/10.1002/9781119183525

Idika, N., & Mathur, A. P. (2007). A Survey of Malware Detection Techniques. SERC Technical Reports.

Vinod, P., Jaipur, R., Laxmi, V., Gaur, M. (2009). Survey on malware detection methods

Jaha, A. A., Shatwan, F. Ben, & Ashibani, M. (2008). Proper Virtual Private Network (VPN) solution. In Proceedings - The 2nd International Conference on Next Generation Mobile Applications, Services, and Technologies, NGMAST 2008. https://doi.org/10.1109/NGMAST.2008.18

Adeyinka, O. (2008). Analysis of problems associated with IPSec VPN Technology. 2008 Canadian Conference on Electrical and Computer Engineering. doi:10.1109/ccece.2008.4564875

Cisco. (2018). What Is a VPN? - Virtual Private Network - Cisco. Retrieved September 28, 2020, from https://www.cisco.com/c/en/us/products/security/vpn-endpoint-security-clients/what-is-vpn.html

Lakbabi, A., Orhanou, G., & El Hajji, S. (2012). VPN IPSEC & SSL technology: Security and management point of view. In International Conference on Next Generation Networks and Services, NGNS (pp. 202–208). https://doi.org/10.1109/NGNS.2012.6656108

Fowler, K. (2016). An Overview of Data Breaches. In Data Breach Preparation and Response. https://doi.org/10.1016/b978-0-12-803451-4.00001-0