TALLINN UNIVERSITY OF TECHNOLOGY

School of Information Technologies

Arefeh Fathollahi Kalkhoran 177241IVCM

A SYSTEMATIC PROCESS TO IMPROVE DATA LOSS PREVENTION IN A LARGE ORGANIZATION

Master Thesis

Supervisor: Hayretdin Bahşi

Prof. Dr.

Tallinn 2021

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Arefeh Fathollahi Kalkhoran

30.11.2020

Abstract

Protecting organizational data from cyber threats is very important from different perspectives. If organizational data including business documents, customer's documents, and employees' documents falls in the wrong hand, fines by regulations and damage to organizational reputation are only two of the consequences of it. Data Loss Prevention contributes to preventing organizational data falling in the wrong hand.

This thesis conducts a process of implementing Data Loss Prevention (DLP) improvement on a large organization[1] with approximately 5550 employees and more than 50 corporate sites [1][2]. The novelty of this thesis is the systematic practical organizational process of implementing a chosen technology related to DLP on a large organization with large amount of data and in a real environment with real interactions of departments and users. This detailed process can be a complete source of guidance for a confused Security Specialist to initiate and proceed confidently through the improvement or first-time implementation of DLP.

Data is generally stored, used and transferred in 3 environments of cloud, on-premise and endpoints in companies. The scope of this thesis includes on-premise and endpoint environments' Data Loss prevention implementation process, leaving the cloud environment relatively for further future study.

An action research study has been applied as the main methodology since there were collaborations among different parties in the organization and the author of this thesis is involved as the technical lead and the owner of the whole implementation of this action research. There will also be various qualitative and quantitative methodologies involved within the process of the implementation.

In relation to endpoint DLP, qualitative interview survey to identify the problems from the end users, qualitative observations on the deployment, and quantitative method to observe the improvement of the DLP are utilized. In relation to on-premise DLP, qualitative observation/explanatory methodology is utilized for the analysis to prove the positive effect of the change made causing the improvement of DLP.

[1] "A large organization has over 500 employees" [2].

Microsoft Information Protection is utilized as one pre-existing product in the subject company. However, the main concentration of this thesis is organizational process through which the accomplishment of a successful and effective DLP is proved to be true.

The result indicates that the proposed process has great deal of effectiveness on the improvement of DLP. The analysis of the documents containing sensitive data and users' manual labeling progress validates a major progress in protecting documents for endpoint environment and optimizing the number of the false positives and the reducing the number of the alerts in on-premise environment resulted in DLP improvement to a certain extent.

This thesis is written in English and is 107 pages long, includes 7 chapters, 26 figures and 4 tables.

Acknowledgements

I would like to acknowledge every person who assisted me in the completion of this thesis. My Parents who Supported me financially and emotionally throughout this curriculum and my best friend who instructed me in applying for master program in Tallinn University of Technology. I also appreciate the assistance of Stefan Sütterlin in the optimization of this thesis's questionnaire structure.

Last but most importantly, I would want to thank the kindest supervisor, Prof. Dr. Hayretdin Bahşi, who patiently and compassionately directed me into this rather large thesis.

List of Abbreviations and Terms

DLP	Data Loss Prevention
MIP	Microsoft Information Protection
IT	Information Technology
DPO	Data Protection Officer
GRC	Governance, Risk and Compliance
GDPR	General Data Protection Regulation
PII	Personally Identifiable Information
OS	Operating System
EU	European Union
US	United States
AIP	Azure Information Protection
SCC	Security and Compliance Center
UL	Unified Labeling
OP	Organizational Process
FISMA	Federal Information Security Management Act
NIST	National Institute of Science and Technology
FIPS	Federal Information Processing Standards
PCI-DSS	Payment Card Industry Data Security Standard
ALM	Application Lifecycle Management
RCM	Revenue Cycle Management

CSO	Chief Security Officer
FDQN	Fully Qualified Domain Name
Vlan	Virtual local area network
IP	Internet Protocol
SQL	Structured Query Language
VM	Virtual Machine
RAM	Random Access Memory
HDD	Hard Disk Drive
vCPU	virtual Central Processor Unit
SCCM	System Center Configuration Management
RegEx	Regular Expression
HR	Human Resource
FP	False Positive
TP	True Positive
TN	True Negative
FN	False Negative
CV	Curriculum Vitae
GPS	Global Positioning System
SOC	Security Operation Center
IRP	Incident Response Plan
MS	Microsoft

IDS	Intrusion Detection System
B2B	Business to Business
B2C	Business to Customer
CIA	Confidentiality, Integrity, Availability

Table of content

List of figures

List of tables

1 Chapter 1: Introduction

1.1 Summary

There is a massive amount of data at rest, in transfer and in use electronically and physically in large organizations. Data can also be transferred between companies, partners, entities and individuals. These mentioned parties may locate in different cities, countries or continents. While data is being stored, used, or transferred electronically, or physically, are we certain that it is shielded appropriately? What if a company's data falls in a wrong hand? One of the most interesting data breaches in 2020 was the vulnerabilities in Zoom which became more and more popular for online meetings due to COVID-19. 500,000 stolen passwords were reported and sold in dark web [3]. There are many cases of data breaches every year and the number of the cases grows day by day.

When valued data becomes discovered, classified and protected in order to prevent misusage of it, data loss prevention is the topic involved. From security perspective, data's value depends directly on the impact of the misusage against the data owner (customer, company, etc.). One of the most important data that must be protected is sensitive data (Includes personal data, health records, and financial data).

There is a huge body of research that is focused on different aspects of data loss prevention. However, the author, as an information security engineer in the thesis's subject organization, identified a gap of a detailed application of the process of data loss prevention improvement and the obstacles, troubles, and blockers throughout the process with a real setting practice. The process includes the phases of the process and the tasks to do for each phase. Improvement in this research means that there has been some minor attempt earlier regarding data loss prevention in the organization, which is identified as both insufficient, and inappropriate from the subject organizations' managerial perspective.

Azure Information Protection (AIP) is a portal from where the protection labels and policies are created and loaded to endpoints' Microsoft office products such as Word and Excel to be applied on the documents. AIP also collects and illustrates logs from the endpoints. Security and Compliance Center (SCC) is another portal that also has the label and policy creation along with providing other features such as retention.

Unified Labeling (UL) feature added to Azure Information Protection (AIP) in connection to Security and Compliance Center (SCC) are together a new sensitive data detection, classification and protection feature provided by Microsoft which is called Microsoft Information Protection (MIP) [4] as a whole. This feature is recently introduced.

Microsoft Information Protection (MIP) process is a rather recent process that will be utilized as an inspiration in the current research. End-user interview, organizational departments involvements, the obstacles and the analysis of the findings will be conducted followed by concluding if the whole process proposed has improved Data Loss Prevention (DLP) status on the subject organization. This process can then become applied on any large organization with much less amount of time and pre-knowledge of the possible limitations and obstacles to be encountered. The obstacles, limitations and the lessons learnt will also be covered in this paper throughout the whole process.

The methodology which is used to best fit the process is mainly action-research methodology. The reason why this method is selected is that there are direct involvement and collaboration between the researcher (the author of this paper) and other participants such as Information Technology (IT) department, Data Protection Officer (DPO), Head of Security department, Governance Risk and Compliance (GRC) department and vendor customer support throughout the whole process. There is a direct involvement of the researcher (This paper's author) in this collaboration. The author is the main participant as the technical lead of the implementation. In some steps, there is also orientation towards a repeated cycle of changes by actions and tests, analysis and researches to remove obstacles and optimize the results.

Both endpoint[1] and on-premise[2] environments will be covered in the scope of this paper. DLP in both of these environments will be conducted in parallel and most of the initial tasks cover both environments. However, half-way through the process, each environment will require their specific steps. The main difference in methodology between these two environments is as follows:

For the endpoint DLP, the methodology of interviews will be conducted to determine the exact difficulties from the perspective of the end-users in using the current document protection labels

[1] Includes physical machines of the corporation (windows 10 and MacOS).

[2] Includes shared file servers owned and maintained internally by the organization.

as well as user behavior log analysis after training the users on using the protection labels. The scope for endpoints is all the windows and Mac machines (5550 in total) in corporate environment.

For on-premise DLP, the methodology of observation was utilized including quantitative log analysis of the discovered files to find the detections' false positives, false negatives, true positives, and true negatives. There will also be considerations on how to reduce alerts made by true positive detections as well as using a condition to reduce false positives. The scope of on-premise DLP will be 1 fileserver[1] in one geographical site. This fileserver contains internal shared files with different purposes such as work-related testing documents, lists of asset information, Human Resource (HR) related documents, memorial event pictures. etc. The scan is conducted on the subject file server to detect sensitive data containing documents. If the scan is successful, this action can be expanded to other file servers in others geographical sites which is out of the scope of this study. In this thesis, 2 scans will be assessed and analyzed.

This thesis's outcome is valuable from both perspectives of science and industry for researchers and practitioners respectively which will be explained in the next section.

1.2 Novelty and research gap

For a security specialist, Information Security engineer, or any other practitioner who intends to start and proceed with DLP improvement of a large organization, there has not been a practical real experienced report or study including possible obstacles, challenges, limitations, and the lessons learnt through the whole process.

The novelty also includes a systematic real environment study supported by action research and reporting the results. Analyzing the interaction between process and technology in real settings is important and cyber security area requires such studies beyond best practices guides who are not giving systematic knowledge and do not analyze the pros and cons of practices.

1.3 Context

The subject organization consists of approximately 5550 employees and 8 major geographical sites, total number of 51 corporate sites and more than total number of 70 production sites. 550

[1] A storage of shared files in an office.

MacOS machines and approximately 5000 Windows 10 machines are identified in the subject organization.

1.4 Scope

The scope includes endpoint and on-premise environments.

- For the endpoints, the scope includes improving DLP across MacOS and Windows 10 machines.
- For on-premise environment, the scope includes a file server that is scanned by the latest version of AIP scanner[1].

1.5 Research questions

The main question of the present thesis is “How can Data Loss prevention become improved in a large organization?”

The sub-questions are as below accordingly:

1. What are the steps to take as organizational process in implementing and improving DLP status on the endpoints?
2. What are the steps to take as organizational process in implementing and improving DLP status on the on-premise file share servers?
3. What are the lessons learnt in the process of improving DLP in a large organization?

1.6 Chapter overview

This paper includes 7 chapters as below:

- Chapter 1- Introduction: This chapter will explain the summary of the paper as well as Novelty, Context, Scope, research questions and chapter overview.

- Chapter 2- Background, Terms and Definitions: This chapter will have a pre-knowledge on all the terms and definitions mentioned throughout the whole paper.

[1] This scanner scans through the files and compares the content of the files against product related predefined sensitive data detection templates and reports the logs to AIP portal.

- Chapter 3- Literature Review: This chapter will consider other studies that are similar or related to this thesis.
- Chapter 4- Methodology: This chapter will explain the methodologies used in each phase of the DLP process.
- Chapter 5- DLP organizational process: This chapter includes process details in both endpoint and on-premise environments that are proposed and implemented by the author and the lessons learnt in different actions.
- Chapter 6- Analysis and Result: In this chapter the main actions will be analyzed and explained based on the practices on the subject large organization. The result related to all the actions taken analyzed regarding both on-premise and endpoint environments will be explained.
- Chapter 7- Conclusion: This chapter will conclude if the process taken in this study have improved DLP status of the subject organization.

2 Chapter 2: Background, Terms and Definitions

2.1 Data Loss Prevention

Data is raw information in any form. Data can be stored in a computer hard drive, on a USB-based device, on a file server, on cloud storage, etc. Accordingly, Data is known to be in 3 main states of at rest, in transfer, and in use [5], [6].

- Data at rest is defined as organizational data that is static and resides on any data storage such as email servers, file share servers, file servers, network attached storages, USB-based devices, personal workstations, etc.
- Data in transfer (motion) is a data that is moving in the network traffic of an organization such as social media related messages, E-mails, etc.
- Data in use is mostly related to the data that is being used by applications and humans for functioning their duties such as a file that is copied by a user or a document created by an application when extracting some list of information from it as an excel file. DLP has two acronyms of Data Loss Prevention and Data Leakage Prevention.

Alkilani et al. come “DLP is a set of technologies, products, and techniques designed to help prevent sensitive information from leaving an organization” [7].

This thesis’s author tends to redefine DLP because the definition above may need some additional explanation. DLP systems and their coverage in terms of file extensions, integrations with other systems are constantly changing and their definitions may change along with them.

The author defines Data Loss Prevention as a solution of protecting classified data while being transferred, stored, or used by either a System, an application or a user on different environments of on cloud, on-premise, and on the endpoint from being exposed intentionally or unintentionally and accessed by an internal or external unauthorized party. However, DLP is not only a technology, but in an organization, there should be a DLP process, policy, etc to be followed to address the proper implementation.

Data Leakage is most of the times referred to as Data breach but there is a slight difference between these two terms. Data leakage is mostly related to a wrong action, intentionally or unintentionally, from the inside of a company (for example an employee) leading to exposing the data to the wrong hand. On the other hand, Data breach is mostly related to an action with which an external person has unauthorized access to an organization’s data (mostly sensitive data).

Sensitive data in General Data Protection Regulation (GDPR) is defined as:

- “personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs;
- trade-union membership;
- genetic data, biometric data processed solely to identify a human being;
- health-related data;
- data concerning a person’s sex life or sexual orientation” [8].

Loss of data refers to a situation in were data is stollen by unauthorized person, modified, or deleted intentionally or unintentionally by an internal or external user. Data breach is a type of data loss which is mostly related to malicious or accidental loss, modification or deletion of Personally Identifiable Information (PII). PII in GDPR is referred to as personal data, and personal data is a type of sensitive data.

Data can be in the form of a simple screenshot, a document in a USB-based device of flash memory, an external hard, or in a document file with different extensions such as txt, xlsx, etc. the important point is that not all the data needs protection. Data, depending on the consequences of its exposure, varies in value and needs to be protected relatively.

Organizations collect personal data from their customers, employees, partners, subsidiaries, and third parties. In case of data breach incident in an organization, one of the many harms is possible fines posed to it. Figure 1 by Statista [9] provides a bar chart of the data breaches reported in the European Union (EU) from May 2018 to January 2020. As the bar chart clearly indicates, the number of data breach cases have increased significantly in all the countries without any exception.

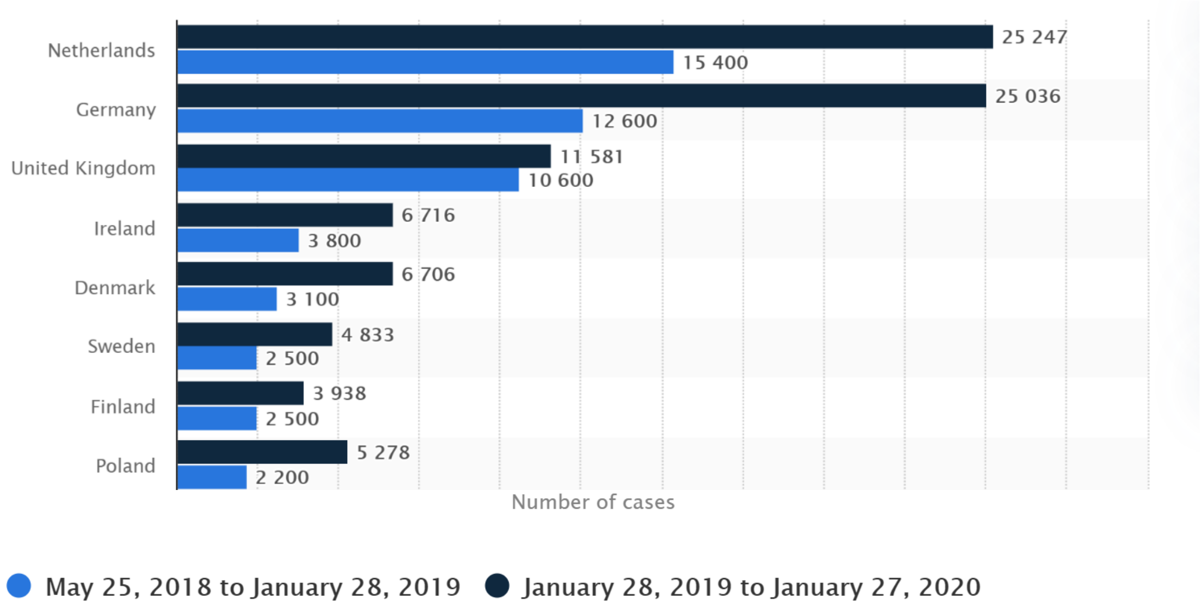


Figure 1. Data breaches reported in the EU from May 2018 to January 2020 [9].

Under GDPR, organizations that fail to apply powerful protection on personal data stored, used, or transferred in it, can be fined up to 2 percent of their previous years revenue if it is their first time of data breach. In case of more than one time, the fine will be increased to 4 percent which is called repeated offence [10].

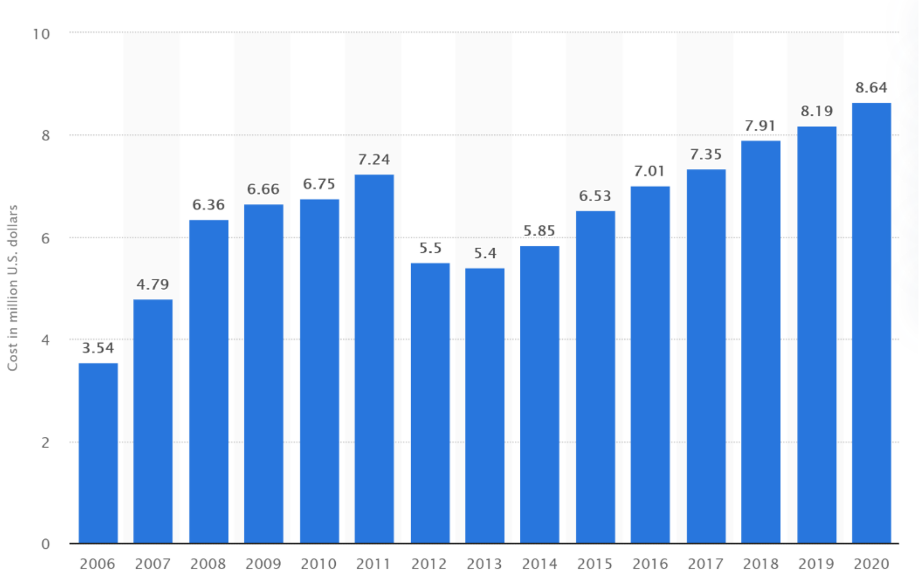


Figure 2. Average data breach cost to businesses in the US [11].

Other than GDPR which is mostly related to EU companies, the US companies also require to protect sensitive data. Statista [11], in Figure 2 also provides a visible diagram of the average data breach cost to businesses in the United States (US). As the chart illustrates, the cost to data breach is growing every year and will most probably grow in the next years if organizations do not prepare themselves with DLP.

Therefore, it is highly crucial that organizations protect specifically sensitive data with the most powerful and recent protection mechanisms to prevent data loss and the consequent fines.

2.2 DLP systems: Microsoft Information Protection (MIP)

"A DLP system includes a set of rules and policies that classify data according to its type to ensure that it is not maliciously or accidentally shared" [7].

There are numerous DLP systems in the market including CoSoSys Endpoint Protector, Symantec Data Loss Prevention, Teramind DLP, Clearswift Adaptive DLP, SecureTrust Data Loss Prevention, Check Point Data Loss Prevention, Digital Guardian Endpoint DLP, Code42, CA Data Protection, and Comodo MyDLP [12].

Microsoft also has provided tools to utilize for the improvement of DLP. Microsoft Information Protection (MIP) [4] is composed of applying classification, labeling, and protection on data. Discovery means to find documents and propose that each document can be in a category based on the information that is in the document. Labeling refers to assigning a categorical name to documents and protection is to configure encryption and user restriction on the labels assigned to the files based on the category that they are.

Label and Policy

After discovery, labels, as mentioned earlier, will be assigned to the documents to categorize or classify them. When the label is assigned to a document, the settings of the label will be applied on the document. The setting parameters include label name, description, color, permissions, visual marking, and conditions. Label name is a name that is assigned to a label and will appear in the end-users' Microsoft (MS) office application. For example, Public, Internal, and confidential are 3 names for labels (See figure 3). Description is a brief explanation of what the label does which will appear when the end-users hover over the labels. Color is a small square beside the label shown to the end-users. The labels can be colored from a light color to a dark color to represent the strength of the label's protection. Permissions will include configuring who should access to what document. Visual marking refers to putting words in the background of the content of document. It can be diagonal word of "confidential" in the background of a document. Conditions include 2 options. To use 1) predefined template or 2) set customized Regular Expression[1] or a keyword to detect files with specific content. Pre-defined templates are the built-in templates that are created based on the regulations around the word regarding sensitive data. If they are configured in a label, they will be automatically applied on the documents that are detected as containing sensitive data. Labels can be set to be applied automatically on a document that contains sensitive data or recommend the end-users to apply a label on the document that is detected as containing sensitive data. Both of auto-labeling and recommending label settings are using the same mechanism of detection by the pre-defined templates [13].

Policy means what to do with the labels. The labels that should appear for users, the way they appear, how many of them appear to whom, where to send the logs, what label should be the default label (if it is needed), what label should be automatically applied (if it is needed), activating options that the users could have such as do not forward button, and customizing user permissions and so on are all configured in Policy [14].

Microsoft has utilized and unified the 2 management portals of Azure Information Protection (AIP) [15] and Office 365 Security and Compliance Center (SCC) [16]. These are the main tools used in this thesis which are explained from page 24. MIP offers a recently promoted hybrid system which includes several components and integrations to best tackle most of the gaps in the market. It is a hybrid system because it needs agent to be installed on windows 10 machines, while it is a built-in feature on MacOS machines.

Hart et al. 2011 [5] concludes that 2 characteristics should be considered in DLP systems including meta data associated with the content, and the fact that not all secret documents in the world are written in English. Both considerations have been tackled on MIP.

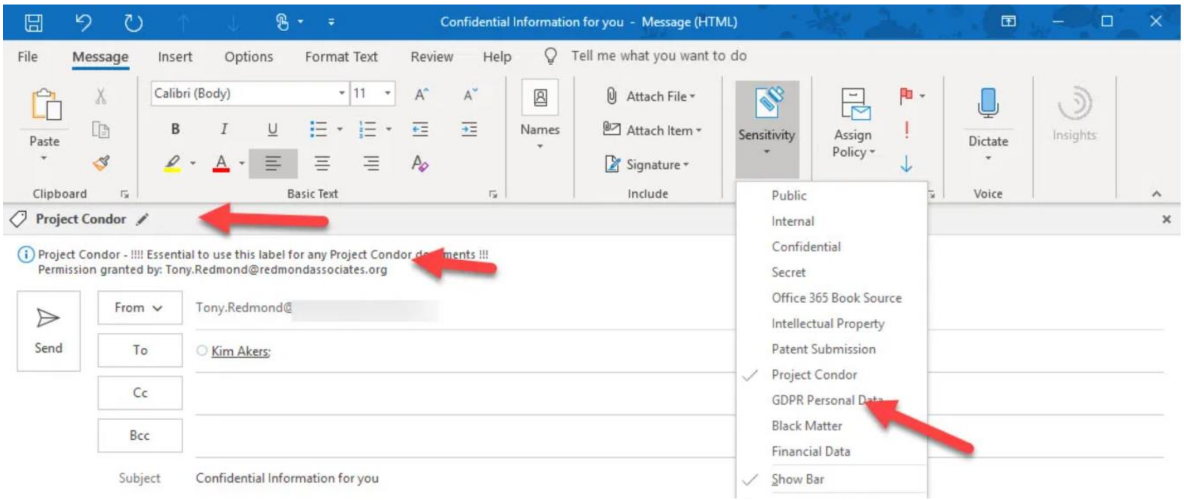
AIP scanner is a software that is installed on top of a windows server and scans the file servers for sensitive data. This scanner has two modes of discovery and protection. Discovery mode

[1] Regular Expression (RegEx) is the regular formation of a string that is repeating in all the instances of the string. For example, the identification card of a specific country starts with 123 and ends with 456 such as 123abcde456. RegEx defines the repeating sections and detects all the strings based on that.

which only needs read permission on the file servers, will crawl through the files and provides information on how many files are in the fileserver, how many files contain sensitive data, what kind of sensitive data is detected in those files and so on. Protection mode which needs both read and write permissions, will then apply labels and policies that are configured by the security practitioner on the discovered files. The labels can be configured to apply encryption or user restriction at this point.

AIP is a management portal that provides the logs. Any log from the Endpoints (collected by the client on windows 10 and built-in feature of MacOS machines) and On-premise (collected by AIP scanner) will be gathered in two blades (vertical tabs) in AIP management portal which are called Data discovery, and Activity log. These two blades will be used for log analysis in this thesis.

AIP also provides the feature of creating a set of labels and policies by the administrators/security specialists. These labels and policies will later be applied manually (by the end-users), or automatically (by the administrator/security specialists for on-premise, and by predefined templates for endpoints) on the documents. The label, which classifies and protects the document, will travel with the document everywhere the document moves or resides.


Confidential Information for you - Message (HTML)

Project Condor

Project Condor - !!!! Essential to use this label for any Project Condor documents !!!
Permission granted by: Tony.Redmond@redmondassociates.org

From: Tony.Redmond@[redacted]

To: Kim Akers;

Subject: Confidential Information for you

Figure 3. Protection labels that are downloaded from SCC to the endpoint Microsoft office outlook [44].

SCC also provides creating labels and policies the same as AIP. Other capabilities of SCC are creating alerts based on the detected sensitive data in documents. The detection will occur based on the settings in the labels and policies. Figure 3 shows an example of labels that are downloaded from SCC on the Microsoft office Outlook.

As a matter of fact, these two portals have some features in common. Each portal also has its own specific feature. Both portals contribute to a stronger protection.

As mentioned earlier, both portals provide setting and configuring labels and policies. To upload the labels to the endpoints, an agent must be installed on the endpoint machines. AIP's agent is called classic client (deprecation as of July 2021) and SCC's agent is called Unified Labeling (UL) client. These two clients are the same product with different versions. UL client is an upgrade to classic client meaning that the two clients cannot tolerate each other on the same machine. There can only be one of the agents installed on a machine. Microsoft has announced that classic client will be deprecated by July 2021 and must be replaced with UL client [17].

UL client will download the labels from SCC and compares files on the endpoint against the labels and policies configured in SCC. SCC and AIP were **not** synchronizing with each other in terms of consistent labels and policies' names and settings in the past.

Microsoft has recently offered Unified Labeling (UL) feature in AIP that provides synchronization with SCC. Synchronizing the names and settings of the labels and policies so that organizations can leverage all the capabilities of both management portals for DLP. UL is a feature in AIP that initiates and continues synchronization of both labels (service is already available) and policies (in preview[1] by the time of this paper) from AIP to SCC.

Other than the benefit of leveraging all the features of both portals, MacOS machines are also only covered by SCC and not AIP. AIP only covers windows 10 machines.

Client (UL or classic), when installed on a device, will appear as a plug-in on all the Microsoft (MS) office products (such as MS word, Excel, etc.). An icon (sensitivity icon in figure 3 is the

[1] Preview means that the feature is functional, but it is in a testing phase and opened for the customers to identify the misfunctions and improve the feature accordingly.

icon for UL client) will appear on MS products toolbar with the labels that are configured on the management portals. Then the users can apply those labels on their documents.

The status of the subject company at the beginning of the DLP improvement related to this thesis was as below:

- Classic client is installed on all the windows 10 machines.
- MacOS machines are not covered.
- AIP has some labels and SCC has some other different set of labels. Both are being used across the whole company which brings manual double work of synchronizing them with any change in the label settings.
- Labels are loaded to the endpoints by classic client from AIP portal on windows 10 machines.

2.3 Action research

Action research is a quantitative methodology with 2 main components of data collection and idea implementation [18], [19]. Action research methodology is collaborative and participatory which is not only guided and acted by the researcher. There are other parties that collaborate with the researcher to complete the implementation of an idea in practice [18], [20], [21], [19]. All the participants together contribute to the result of the implementation of the idea.

Other than collaboration which is one of the characteristics of action research methodology, it is empirical, and it attempts to solve a real-world problem. In parallel of solving a real-world problem, the experience through the whole project/research is recorded which is of value [22], [23].

In summary, most action research studies have four characteristics in common: “(1) An action and change orientation (2) A problem focus (3) An “organic” process involving systematic and some- times iterative stages (4) Collaboration among participants” [24], [23].

Tüzün et al. 1999 [24], [23] (Figure 4) mentions that action research consists of five basic steps:

1. Diagnosis: A problem in the organization becomes identified.
2. Action planning: The actions to take to address the problem will be planned.
3. Action taking: The actions that have been planned will be taken.
4. Evaluating: The influence of the actions taken will be measured.

5. Specify the learning: lessons learned throughout the whole steps above will be recorded.

The steps above will be utilized in this study. Step 2 to 4 may repeat if the evaluation does not show that the problem is solved.

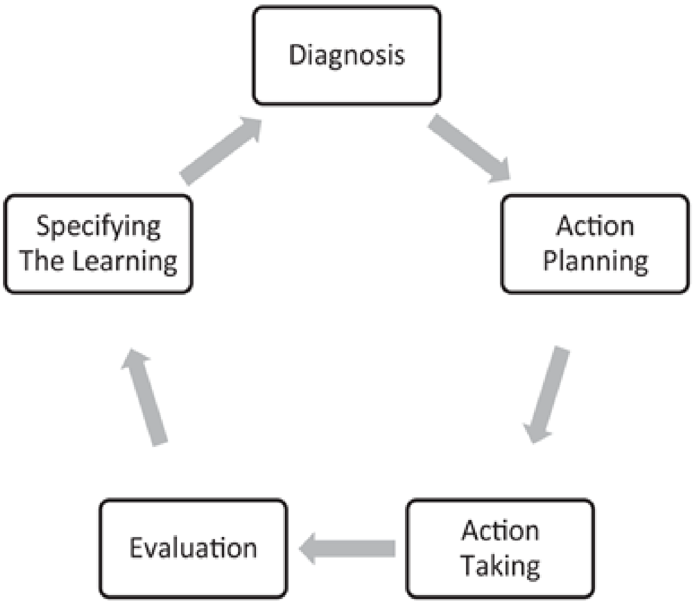


Figure 4. basic steps of action research methodology [21].

3 Chapter 3: Literature Review

3.1 Organizational process

This thesis follows and focuses mostly on the organizational process of implementing a technology. Organizational process as Bamel et al. claim is defined as "the method adopted/followed by an organization to achieve its goals" [25].

As Magdaleno et al. claim, "It is almost impossible to define a standard process due to varied and unpredictable cases encompassing while conducting a project [26]. This means that a single process for all the projects in an organization in different fields is impossible. However, in one field of study, a single process, can be utilized. This thesis is specifically designing a process for all DLP deployment of document protection in large organizations. The Process for software development is shown in Figure 5 [26]:

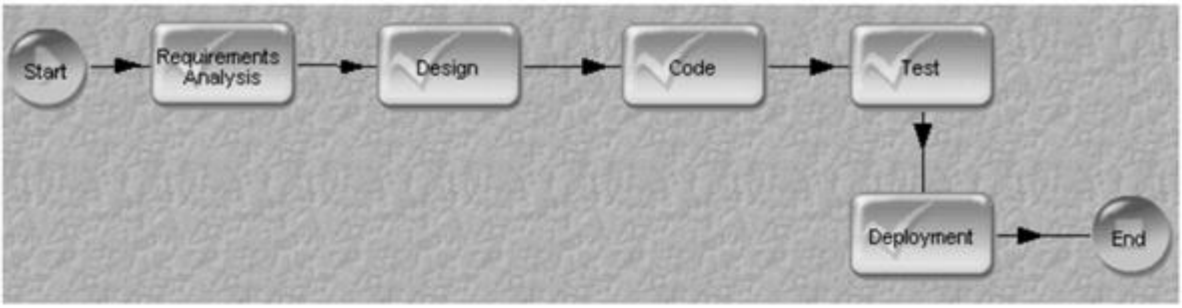


Figure 5. Process for software development action research [20].

Similar Cycle has been introduced by Tüzün et al. [23] related to the lifecycle management within a large-scale company, which is very similar to what has been seen in Kumar et al. study earlier discussed [25] (see Figure 6).

This thesis will also come up with process phases of DLP that can be applicable on all the organizations. A set of phases are needed for any process. After that, there is a need to propose tasks within each phase which in this thesis will be proposed by the author.

Magdaleno et al. divided the phases into beans and called them process beans. They introduced these beans as components that complete each other to achieve the purpose. The process beans that were introduced for software engineering as an example throughout that research are Analysis, Design, Code, Test or Deployment [26].

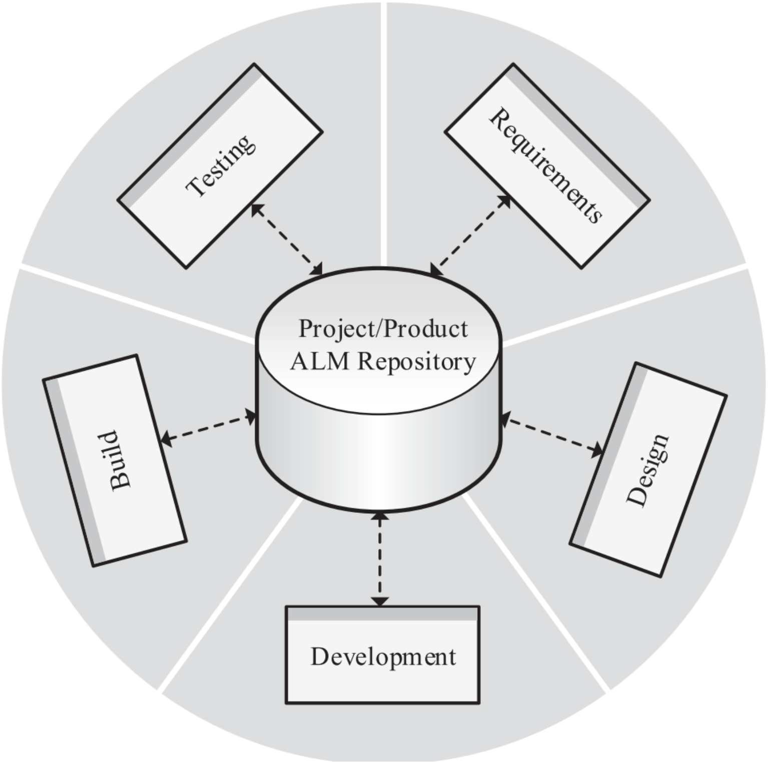


Figure 6. Action research lifecycle [21].

although these process beans will not fully meet the approach of DLP implementation, they are not irrelevant. Planning, testing and deploying are repeatedly used as process phases throughout this thesis. In this thesis, the phases of discovery, classification and protection are inspired by the proposed-by-vendor general phases of DLP implementation. This organizational process will be a guiding framework for this thesis [27].

This thesis's author also identified a point that there is a connection between action research methodology and Organizational Process (OP). Kumar et al. claim that "… teamwork, communication and collaborative decision-making, workplace for creativity and performance management system were taken as sub-dimensions of OP" [25]. It has been mentioned earlier that one of the characteristics of action research methodology is collaborations between participants. Therefore, it seems that organizational processes are mostly accompanied by action research methodology which is a good reason why this thesis is based on action research methodology.

3.2 Data Loss Prevention (DLP)

Data Loss Preventions seems to have a long history of importance. It came into use in 2006 and became popular in 2007 [28], [7]. So many researchers have conducted research on different aspects of Data Loss Prevention such as data exfiltration, risk reduction, and benchmarking DLP systems, etc. However, recent DLP related researches are mostly focused on cloud environment which is more contemporary technology and has more immaturity than other environments of endpoint and on-premise. While organizations are turning more to cloud environment, they still have data stored and used on the endpoints and on-premise which need to be protected comparatively with the latest technology related to them.

There are numerous tools that are being used in different departments of a large organizations with branches all over the world. These tools can be created by the company itself, supported by a vendor but maintained by the company itself, or supported and maintained by a vendor. Assume that one piece of data can be passed through these tools to and from third parties by numerous integrations between applications and tools in different environments. How can data become protected while going everywhere?

Data must be protected, and the protection is necessary to stand by the data as metadata wherever the data moves to or resides. There are several solutions in the market that provide Data Loss Prevention. However, different products have advantages and downsides.

A benchmarking research by Glen in 2015 [29] was conducted on 4 DLP systems to evaluate their operations, and abilities. These 4 systems include My Endpoint Protector [30]: Trustwave [31], MyDL [32] and OpenDLP [33].

In that research OpenDLP was immediately considered to fail to cover data in motion and data in use. It could only cover data at rest. Also, Tustwave had issues with improperness of notifying the user in a misleading way when the user attached a file with sensitive data to an email [29]. In that research, DLP systems are divided to agentless, agent-based, and hybrid based on if there is/is not a prerequisite of an agent installation on the endpoint for them to function. DLP in MIP in this thesis is a hybrid that uses agent on the windows 10 machines and is agentless[1] (built-in) in Mac OS workstations.

Alkilani et al. [7] study is one of the most recent studies related to DLP which reviews on data exfiltration techniques as well as attempting to bypass DLP systems. Raman et al. [6] Focuses on the problems related to DLP technology solutions such as signature misuse detection and lack of detecting complex data loss scenarios. It seems like most of DLP technologies compete in 2 main protection capabilities which are encryption and access control.

Alkilani et al. [7] mostly focus on comparing network protection technologies' mechanisms referred to as firewalls, and Intrusion Detection System (IDS) in the network with DLP systems. In that study it was predicted that DLP technology would progress to cover smartphones in 2014 which is now an available technology of recent DLP services such as SCC and AIP. Lopez et al. [34] evaluate DLP systems capabilities and technical features. Praba et al. [35] introduced DLP system capabilities such as encryption, user restriction/access control and policies. It also identifies the main difficulties in the DLP systems from an administrator perspective. For example, user access rights are challenging. A document can be protected to restrict becoming printed while once the receiver needs to print it based on their own internal requirements of stamping the document and scanning it and sending it back. This is what is mostly referred to availability problem. These capabilities are also what AIP and SCC offer. Three phases in Alkilani et al. study for DLP implementation were utilized as below [35]:

1. Data collection
2. Data analysis
3. Remedial action

These phases are very similar to the 2 phases utilized in this thesis which are Data Classification and Data Protection.

Alkilani et al. [7] conducted a series of scenarios to test the capabilities of Symantec DLP. Data exfiltration techniques were tested mostly on MS word and Adope's PDF files to find the weaknesses of the product by attempts to bypass the DLP system. Symantec DLP, similar to AIP which is used in this study, includes a set of rules and policies to detect files by comparing

[1] Agentless refers to the fact that there is no need for the Mac machines to have the agent installed on them as they already have the feature within their operating system.

against Regular Expressions (RegEx), keywords, and patterns. Alkilani et al. tested the DLP system, Symantec DLP, by modifying the file content and extension.

The author of this thesis also tested AIP similarly at the initiation stage of the implementation as part of the very first phase. Another similarity is when Alkilani et al. [7] mentions that there is a default feature of pre-defined built-in rules and policies which can be used on demand. These rules and policies can be customized by the administrator. Even though as part of this thesis the product to be used throughout the study is also tested for the purpose of evaluation, the product is not the main focus of this thesis. AIP which is a part of a related license was already purchased by the subject organization and testing the product was conducted only on the purpose to decide if AIP is worth dedicating time and human resource for DLP improvement. This is because AIP feature was already a part of a purchased license in the subject organization.

Kaur et al. [36] explains the improvements that needs to be considered by DLP systems' service providers regarding access controls, encryptions, etc. Polozova proposes a threat model while emphasizing the risks related to malicious insiders attempting to bypass the DLP systems.

3.3 Data Classification

The segregation of an organization's data into different categories based on the value of the data for the organization is called data classification [37]. It is also important to mention that a piece of data can be valuable for one organization, while considered not to be so valuable for another organization.

Hart et al. claims that "Research in the document classification field dates back to 1960s" [5]. The author of this thesis discovered that data classification idea even goes further back to 1957 [38]. This study was not classifying documents for the purpose of protection though. It had an attempt to classify words to address searching the literary information. Hart et al. [5] stepped ahead and created their own algorithm of less than 3% of false positives which is a very low rate.

In this thesis, the formula introduced in Micheal's paper will not be used exactly to evaluate the accuracy of the findings in both endpoint and on-premise files detected. However, due to the massive number of files detected as containing sensitive data, a scope will be chosen in both environments and the content will be checked manually (for on-prem, content of files is accessible but for endpoints the author needs to check the activity log) to identify False Positives (FP) and calculations will not be for the whole findings. “lower false discovery rate (FDR), i.e., the percentage of false alarms depends on True Positives (TP) and is defined as:

FDR = FP TP +FP” [5]

Rajagopal et al. [37] introduced data classification steps based on the regulations and frameworks mostly related to governmental organizations in America. National Institute of Science and Technology (NIST) 800-60 framework as well as Federal Information Security Management Act (FISMA) and Federal Information Processing Standards (FIPS) are used to conduct the steps of data classification in that study. This thesis is based on the internal policies of the subject company located in Europe. These internal policies are conducted based on the regulations that apply to EU companies such as Payment Card Industry Data Security Standard (PCI-DSS), General Data Protection Regulation (GDPR), etc.

Nevertheless, data classification steps are more or less similar globally. Recent technology steps further in covering some actions expected from the user to be done automatically. However, it is also important to mention that a very common question may be raised by any participant in the action research process that: Why not protecting all the data that we have regardless of their importance? A security specialist needs to persuade the participant to lead their mind to the better approach. This scenario occurred when the author intended to discuss security of a huge important database of the subject company with the top database manager.

The answer to this question is that in large organizations massive number of files containing data become transferred, created, printed, copied, saved, edited, etc. Moreover, there is always concerns about tradeoff between security(confidentiality) and availability. When a document becomes protected because it contains sensitive data, protection mechanisms such as encryption and user restriction will be applied on the file. Reading the document can be allowed for an external receiver but other actions (copy, print, etc.) restricted. Assuming that we use this single mechanism for all the data of the company would be a disaster of resource consumption and availability problems for the not-valued data to be edited, printed and copied by users other than the owner (creator of the file).

Rajagopal et al. [37] also has similar belief and mentions that:

“Data classification takes this to an importance-based paradigm, where data that is more important is better protected - the homogeneous approach provides less protection than necessary for high-value data and more protection than necessary for low-value data. This leads to more resources being used on over-protecting trivial data which can otherwise be used for protecting important .” [37].

Rajagopal et al. [37] also bases classification on the organizational impact of data if exposed. These levels are inspired by FIPS as low, moderate and high imposing limited adverse, serious, and severe effect respectively. Data classification steps may vary depending on urgency, product used, organizational requirements and limitations. MIP product that is used in this research will cover almost all the steps introduced by Rajagopal et al. automatically. (Figure 7)

In Figure 7, discovery mode will be used to identify sensitive data for on-premise environment. Security objectives (confidentiality, availability and integrity) are not criteria used for classifying information. The existing internal policy is based on impact. Protection mode will be the application of security countermeasure after discovery. The last phase will not be covered in this thesis since all the actions will be taken automatically for the current moment and the future and there is no need for a guide.

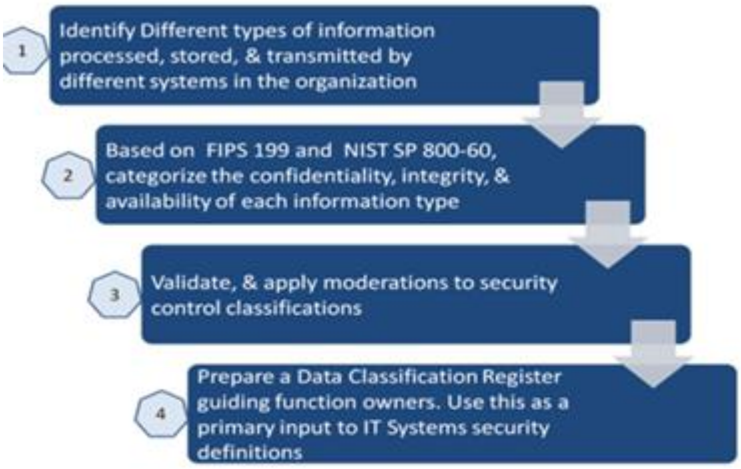

1
Identify Different types of information processed, stored, & transmitted by different systems in the organization
2
Based on FIPS 199 and NIST SP 800-60, categorize the confidentiality, integrity, & availability of each information type
3
Validate, & apply moderations to security control classifications
4
Prepare a Data Classification Register guiding function owners. Use this as a primary input to IT Systems security definitions

Figure 7. Data classification steps [35].

A case study framework by Alparslan et al. was recommended to classify internal documents of an institute using classification algorithms that inspect files for the frequency of the appearance of words to classify the internal documents [39].

Data classification in this thesis will be based on applying labels on the documents that indicates their importance. Moghaddam et al. [40] refer to labels as Data Classification index in 2014. This index was based on Confidentiality, Integrity, and Availability (CIA) as parameters identifying the importance of data. Specific focus of that paper was on the cloud environment which is more or less the similar approach for other instances.

3.4 Action research

This thesis's author has discovered that action research has been widely used as a research approach in social science, and it has been gradually adopted for information systems and software engineering research during the last two decades. Most of the action research studies are related to how technology help improve different industrial contexts such as hospitals, and software development companies.

The study by Tüzün et al. [23] discovered to be one of the most relevant studies to this thesis. Tüzün et al. evaluates the impact of adopting Application Life cycle Management (ALM) on a large organization. ALM is a paradigm for integrations and management of the actions throughout software development and maintenance [23]. Tüzün et al. identified the benefits of ALM and also filled a gap that was mentioned as "Unfortunately, action research papers that include all the steps are still lacking" [23].

The list below includes the similarities discovered between the Tüzün et al.'s study and this thesis's study:

- Utilizing action research as the methodology.
- Application on a real industrial context.
- Obstacles, benefits and lessons learnt.
- Limitations in large organizations.
- Applying a change in the organization.
- Utilizing some tools throughout the study.

There is also the most important similarity related to the novelty of the current research and the research done by Tüzün et al. Tüzün et al. claims that "The results are not only important for the company in which the action research has been carried out but also provide important general lessons for similar companies. The derived lessons are in the first place related to the topic of the action research, that is, the adoption of ALM" [23].

On the other hand, there is a difference involved:

- The tool used in the this thesis is MIP set of tools. However, the product is not the focus of this study. This product is only used to deploy the DLP improvement. Whereas, Tüzün et al.'s study is mostly product oriented focused on the ALM product.

A typical Action-research cycle is introduced by Tüzün et al. as in figure 4 illustrated earlier. This cycle will be utilized throughout this study. Some of the phases will be repeatedly done which will be added to the lifecycle by this thesis's author and will be illustrated as additionally to the cycle in the figure 4. In other words, when a change is planned (action planning) and applied (action taking), in case of a problem caused by the change, another action will be taken to address the problem. This action can be a roll back or change in the label or policy configuration in this thesis.

Rajendra et al. [41] has studied on the role of IT in addressing the information sharing and coordination challenges related to Revenue Cycle Management (RCM) in hospitals. RCM is a system that manages all the healthcare services from patient registration to payments. Another example is a study by Olesen et al. [42] who conducted a groupware product named Lotus NotesTM in order to facilitate communication and collaboration among the senior management of an institute.

St-Pierre et al. [43] presents an action research study on implementing information technologies on a site for home-made applications. In this article, home-made application is not clearly defined and was only mentioned twice throughout the study. However, the context shows creating a technology on a scope of an organization.

4 Chapter 4: Methodology

As explained briefly in 2.3, the whole thesis is conducted based on action research. For the phases related to each environment, combination of qualitative methodologies such as observation, exploratory, and interview will be utilized and for the analysis of the findings, there will be some quantitative data collection. In the present chapter, all the methodologies used in DLP improvement of two environments of endpoint and on-premise will be explained.

Action research methodology has the characteristics below:

"(1) An action and change orientation (2) A problem focus (3) An "organic" process involving systematic and sometimes iterative stages (4) Collaboration among participants" [24], [23].

The reason why this methodology is applicable on this thesis is the fact that the characteristics above apply on the characteristics of this thesis as below:

(1) An action and change orientation: In this thesis, actions will be taken in order to accomplish an improved result in DLP. This action can be a change in the label taxonomy, a training, etc.

(2) A problem focus: The problems identified specifically for each company can differ depending on the current status. For this thesis, the problems to address throughout the thesis and by the actions taken are explained in detail in chapter 6.

(3) An "organic" process involving systematic and sometimes iterative stages: The fact that DLP improvement is being applied on a real organization with real user interactions, is one of the novelties related to this thesis. It also is presenting the process to take for DLP implementation which is another reason why this study is an action research study.

(4) Collaboration among participants: In this thesis, the management of the whole process from planning down to implementation and protection is conducted by the author. However, in large organizations, the most important point is that every single person has defined responsibilities and tasks. Therefore, some other team members or other teams collaborate in the completion of some tasks. For instance, the label taxonomy can be planned by the author, but to publish them across the whole company, the labels must be approved by some managers such as Chief Security Officer (CSO) or head of security department. The participants include but are not limited to the list below:

- IT: Technical installations and configurations
- Data Protection Officer (DPO): Consulting about different issues
- Head of security: Approvals
- Vendor (Microsoft): Troubleshooting technical issues
- Security members: Testing purposes

The process proposed by the author starts with the phases, followed by the tasks related to each phase. All these phases and tasks will be explained in detail in chapter 5. The rest of this chapter will explain the methods used for each phase related to each environment of endpoint and on-premise.

Table 1 illustrates each phase and the analysis methodology that will be used to cover the phases related to endpoint and Table 2 for that of on-premise DLP process. These tables also contain basic steps of action research methodology introduced by Tüzün et al. In other words, action research phases and tasks are all considered together.

The only exception is that the last phase (referred to as basic step by Tüzün et al.) of action research methodology, specifying the learning, will be evolved throughout the whole process instead of being the last basic step. This means that the lessons learnt will be mentioned for each basic step of action research and will not be mentioned as one single basic step at the latest stage.

4.1 Endpoint DLP

4.1.1 Process methodology

In the first phase, in addition to researching through the product to be used and the current status of DLP in the subject organization, a survey was conducted to identify the DLP problems in the subject organization. This survey is also used as the basis for the actions to be taken. All the phases as well as the actions related to them for endpoint DLP are shown in Table 1.

Table 1. Endpoint DLP action research process.

Tüzün et al.’s Action research Basic steps [23]	**Action research phases for DLP**	**Action**	**Methodology**
Diagnosis/Evaluating	Initiation	Questionnaire Recognitions	Qualitative-Interview
Action Plan/Evaluating	Testing / Planning	Endpoint initial tests And planning the phases	Qualitative-Observation/Exploratory
Action taking/Evaluating	Preparation	Installation of the clients	Quantitative
Action taking /Evaluating	Data discovery on endpoints	Endpoints recognition	Qualitative
Action taking/Evaluating	Classification and labeling	Label taxonomy creation and Training the staff	Quantitative log analysis
Action taking/Evaluating	Applying protection based on sensitivity	Set protection on labels	Quantitative
	Monitoring	Not covered in this thesis	Not covered in this thesis

Diagnosis/initiation

A structured questionnaire was conducted and targeted on 2 geographical sites of the subject organization. Target population for the questionnaire survey is all the 637 employees in total in both sites. The questionnaire includes 11 questions in total with branching options in Microsoft forms platform (Appendix 1).

The author proposes searches for 10 categories of information in the questionnaire to plan the actions accordingly.

1. Identifying The departments that are directly in contact with sensitive data.
2. The number of the employees who are aware of protection labels out of total respondents.
3. The number of the employees who use the protection labels on sensitive data containing documents.
4. The availability limitation’s level of extension of protection labels from user’s experience.[1]

[1] The availability limitation extension means that if employee 1 chooses a label for their document and sends/shares the document to/with employee 2, and the label applies only read permission on the file, employee 2 cannot edit the file and has to ask employee 1 to change the label or permission which takes time, confuses the employees and brings availability problems.

5. Deciding if the employees need training for using the labels.
6. The number of the employees who are aware of internal related policies.
7. The complains/concerns of the employees about the current status of the labels.
8. The suggestions of employees.
9. Deciding if it is better to have a new set of label taxonomy.
10. Determining if there is a need to have default label, automatic labeling assignment, or label recommendation prompting.

The questionnaire will help result in what variables of label configuration and taxonomy should be changed and towards what value. To check if the changes applied has met the hypothesis (improvement of DLP), an alternative method of behavior log analysis using the Activity log in AIP was utilized.

The author collected solutions to problems in the subject organization as below:

1. Documents become labeled automatically or automatically recommend the users to label their sensitive data containing documents.
2. The author proposed a set of label taxonomy and defined a default label. The author proposes that the documents should at the very least have one label applied which is neither very highly protective, nor very lowly protective.
3. Label descriptions need to be changed to those of more comprehensive (easy to understand for the end-users).
4. The number of the labels are as small as possible.
5. Employee training for using the labels.

Proof: Evaluating user behavior before and after variable changes by checking logs. This means how much the users are more interested, comfortable and knowledgeable to use the labels. Activity log that is used for this analysis is illustrated in Figure 8. As in Figure 8, the filtering options include the Activity date which allows the author to take all the actions taken by the end-users on documents in a limited amount of time. Other filtering options are label, Internet Protocol (IP) address, application name, etc which are illustrated in Figure 8 in more details.

Variable (**defined as attribute of an object of study)**: configurations of the labels and policies, user-friendliness, DLP progress/regress, user knowledge. The variables are Categorical type binary since the configurations are mostly choosing an option, or not choosing it (Yes/No).

The questionnaire was sent to the participants via email and closed in 2 weeks. A reminder was also sent 1 day before the closure of the questionnaire. The content of the email is in Appendix 3-4 and the questionnaire is in Appendix 1.

Action Plan- Testing/Planning

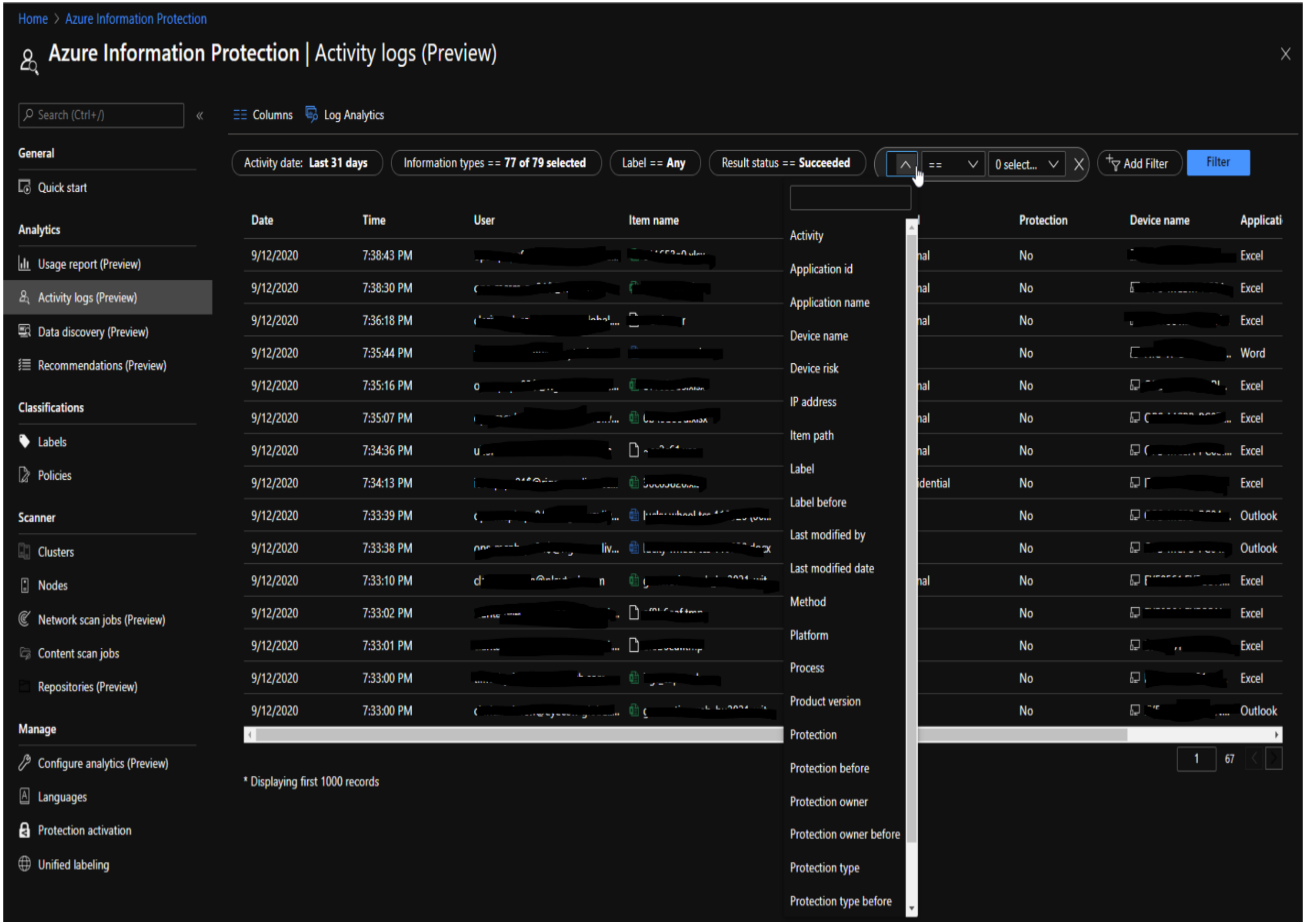
The result of the questionnaire, 10 categories of information mentioned earlier, will be the basis for planning the change. In a large organization, any change that could cause confusion or a sudden tension or panic to the end user side must be tested first on a small group before practical extension of the change across the entire organization. As a part of the change planning, a group test of the security team of 3 people was chosen to test the capabilities of the existed product one of whom owned a MacOS and two of whom owned Windows 10 machines.

The plans and tests were conducted from the AIP management portal. The labels were created, the configurations were changed, and the result of the tests which were observable from the activity log were analyzed to determine the properness of the product functionality. From the end-use's side, the visible changes were also recorded to determine accuracy.

The tests were as below in summary. More details of the tests will be explained in chapter 5.

1. basic checks
2. auto-labeling- custom condition
3. auto-labeling- template condition- label recommendation
4. auto-labeling- template condition- label enforcement
5. label load to SharePoint online
6. label load to OneDrive for business
7. organizational specific data

There has been also a training on how to use the labels and for what types of documents what label is more appropriate. This mandatory training was be a part of the annual information security training which was published to all the employees of the organization on all the geographical sites.



Home > Azure Information Protection

Azure Information Protection | Activity logs (Preview)

Search (Ctrl+/)

General

- Quick start

Analytics

- Usage report (Preview)
- Activity logs (Preview)
- Data discovery (Preview)
- Recommendations (Preview)

Classifications

- Labels
- Policies

Scanner

- Clusters
- Nodes
- Network scan jobs (Preview)
- Content scan jobs
- Repositories (Preview)

Manage

- Configure analytics (Preview)
- Languages
- Protection activation
- Unified labeling

Columns Log Analytics

Activity date: **Last 31 days** | Information types == **77 of 79 selected** | Label == **Any** | Result status == **Succeeded** | == | 0 select... | Add Filter | Filter

Date	Time	User	Item name		Protection	Device name	Applicati
9/12/2020	7:38:43 PM	[redacted]	[redacted]	nal	No	[redacted]	Excel
9/12/2020	7:38:30 PM	[redacted]	[redacted]	nal	No	[redacted]	Excel
9/12/2020	7:36:18 PM	[redacted]	[redacted]	nal	No	[redacted]	Excel
9/12/2020	7:35:44 PM	[redacted]	[redacted]		No	[redacted]	Word
9/12/2020	7:35:16 PM	[redacted]	[redacted]	nal	No	[redacted]	Excel
9/12/2020	7:35:07 PM	[redacted]	[redacted]	nal	No	[redacted]	Excel
9/12/2020	7:34:36 PM	[redacted]	[redacted]	nal	No	[redacted]	Excel
9/12/2020	7:34:13 PM	[redacted]	[redacted]	idential	No	[redacted]	Excel
9/12/2020	7:33:39 PM	[redacted]	[redacted]		No	[redacted]	Outlook
9/12/2020	7:33:38 PM	[redacted]	[redacted]		No	[redacted]	Outlook
9/12/2020	7:33:10 PM	[redacted]	[redacted]	nal	No	[redacted]	Excel
9/12/2020	7:33:02 PM	[redacted]	[redacted]		No	[redacted]	Excel
9/12/2020	7:33:01 PM	[redacted]	[redacted]		No	[redacted]	Excel
9/12/2020	7:33:00 PM	[redacted]	[redacted]		No	[redacted]	Excel
9/12/2020	7:33:00 PM	[redacted]	[redacted]		No	[redacted]	Outlook

1 67

* Displaying first 1000 records

Figure 8. Filter options in Azure Information Protection.

For confidentiality reasons, the training content and the method of training that was used will not be illustrated in thesis. However, it is important to mention that for any organization, the points below must be considered in the training.

- It is better for the training to be a document so that the users find any information in it as fast as possible with a simple Ctrl+F. But still it is good to ask their preference in the questionnaire.
- It is preferred to be short about one page if it is a document and about 3-5 minutes if it is a video.
- It must mention what types of documents must be labeled with what label name.
- It must mention the importance of document labeling.

Action taking- Preparation: The next phase for the endpoint DLP is to take actions. The Unified Labeling client will be requested by security and installed on the endpoints by IT department. At this stage, the change in the configuration variables will be applied and published across the organization. A training platform can also be conducted and shared across the organization.

Action taking- Discovery: The tools utilized to have visibility on the installation progress were Service Now Inventory and System Center Configuration Management (SCCM). Searching and filtering by the name of the applications installed on the endpoints are some of the functionalities by Service Now Inventory tool. The methodology is identified as quantitative since the number of the endpoints that have the UL client installed will be measured and compared against the organizational success criteria. The discovery refers to the fact that Endpoints, as well as the Activity logs are visible.

Action taking- Classification: Activity log was used to determine if there was a change towards DLP improvement in the activities by the end users.

Action taking- Protection: Automatic or recommending labeling will be analyzed to identify if they are beneficial in the improvement of DLP in the subject organization. Depending on the managerial preference, one of the options of automatic or recommending labeling will be deployed.

Monitoring: This phase will not be covered in this paper**.** At the ending stage, corrective actions should be taken in case of any problems identified from the end-user's side and modifications in the configurations of the labels and policies were required. For instance, a user intending to send some sensitive data to an external new subsidiary and the subsidiary is not able to see the content of the sensitive containing email. The reason is that the domain of the receiver's email address has not been added in the configuration of the label. The change needs to be made by adding the receiver's domain to the legitimate domains in the label configuration.

4.1.2 Metrics

The metric to evaluate the endpoint DLP improvement success is the number of the files that are protected after the change in the label taxonomy. After identification of the problems, the action of designing and changing the labels will be taken. If there is a need, training the end-users on applying the labels on their documents, the activity log which collects logs from the endpoints, can be analyzed to measure the actions that are taken by the end-users.

4.2 On-premise DLP

4.2.1 Process methodology

Diagnosis- Initiation: This basic step is qualitative since it involves research through the product, the current status of the organization, the problems and the expectations. At this basic step, the main problem must be identified first. In order to identify the problems, the author contacted different persons in both security department and IT department. GRC specialists as well as DPO and the security managers were also contacted, and information was gathered accordingly.

Table 2. On-premise DLP action research process.

Tüzün et al.'s Action research basic steps [23]	**Action research phases for DLP**	**Action**	**Methodology**
Diagnosis	Initiation	Recognition	qualitative
Action Planning/ Evaluation	planning	Organizational checks	qualitative
Action taking/ Evaluation	Preparation	Installation	qualitative
Action taking/ Evaluation	Data discovery on File Servers	Scanning	qualitative observation/exploratory
Action taking/ Evaluation	Data classification	Log analysis	qualitative observation/exploratory
Action taking/Evaluation	Data Protection	Applying protection on files.	quantitative
	Monitoring	Not covered in this thesis	Not covered in this thesis

Action Planning- Planning: This step is also qualitative since it involves researching. Generally, information about on-premise, SharePoint and SMB file shares must be gathered. Each IT department related to each geographical site must be contacted and requested information Fully Qualified Domain Name (FDQN) or IP address of the file server. In case of intention to implement the scanner first on only one file server, a site must be chosen, IT focal points from that site must be contacted and asked for information on a not critical file server for first scans purpose. As for the scope of this thesis, the author contacted one site's IT managers to gather information about the targeted file server.

Other information such as system requirements, a user with appropriate permissions, optimization and preparations for the scanner installation was gathered from the vendor's knowledgebase. Expectation from the product and its implications are identified at this step.

Action taking- Preparation: When the information is gathered about the target file servers for the scan, the author requested the installation of a windows server machine as well as a user creation on the host windows server with at least read permissions to the targeted file server. Read permission is required for discovery mode only which means that the scanner will only search for sensitive data and returns the logs of all the files including sensitive-data-containing files only. No protection is involved yet. The IP address of the host windows server can either be requested or applied by IT depending on the procedures of the company. In case of a small company, the scanner is best to be in the same Virtual local area network (Vlan) of the file server which contributes to less scan duration. However, in large organizations, it depends on the management decisions. A local/remote database should also be requested from IT department.

Other considerations are the Firewall rules to be requested. It includes opening the connections that are needed between the windows server machine/virtual machine, the management portals, and the target file servers as well as the database for storing the scanner configuration.

This step is rather qualitative since some technical preparations, and repeated troubleshooting actions will be managed. Since the author proceeded with some actions to gain experience in the whole deployment process, it can fall into exploratory and observations. After the plans, some preparations must be conducted for the scanner installation which requires the highest amount of cooperation between different teams. For any installation of scanner for data discovery there must be some preparation actions to be taken. For the specific case of this paper, on the AIP management portal the actions below must be taken.

scanner installation steps to take [44]:

- windows server machine installation by IT.
- Installing patches on the windows server machine by the author.
- Firewall requested by the author and created by other security specialists.
- Proxy on the windows server machine set by the author.
- Static IP net connection requested by the author and set by IT department.
- Setting a content scan job on AIP portal by the author.
- Assigning repository path to the content scan job set by the author.
- Install Structured Query Language (SQL) server requested by the author and installed by IT.
- Conducting test connectivity from the windows server machine to the internet (or at least to the portals), file server and the SQL server done by the author.
- Download and install client AzinfoProteciton_ul.exe on the windows server host.
- Creating a profile, content scan job done by the author.
- Update and Restart the scanner server.
- User permission check done by the author.
- Taking a snapshot from the Scanner server Virtual Machine (VM) requested by the author and done by IT department.

- From PowerShell installing scanner (needs sql server instance path)

Install -AIPScanner -SqlServerInstance <the name of the SQL server> -profile <the name of the profile> done by the author.

- Verifying the installation by the author.
- Authentication maintaining: one Azure Active Directory token for the scanner (none-interactive scan running). App registration in Azure portal for automatic sign in from the scanner to Azure portal.
- A node should appear in the AIP portal after authentication.
- The logs should appear in event viewer in scanner server.

Action taking- discovery and classification: After all these steps, preparations are all set for the first scan on the target file server. The scan can be initiated from both the AIP management portal, and the PowerShell command from the windows server machine. There will be 2 discovery scans conducted and the result will be analyzed.

The condition which will be used reduce false positives and alerts are confidence and count. Confident is the percentage of the certainty that the scanner has detected the sensitive data correctly. If the confidence number is high, the scanner is more certain that the sensitive data was detected correctly. Count number refers to the number of the sensitive data that has been detected in a file such as 10 credit card numbers detected in a file. These thresholds are determined by the author and consulted with the managers.

The false positives will be identified by the logs as well as checking the content of the documents that are in the file server with the account that had read permission which was explained earlier. False negative detections will be calculated by setting a ground truth data set.

Action taking- Protection: At this point the scanner mode should be changed from discovery to protection. Then the preferred settings such as creating a policy and alert will be configured.

Eventually protection labels will be applied on the sensitive containing files in the file servers on the files that were detected as true positive in containing sensitive data in a quantitative methodology.

Monitoring: This phase will not be covered in this paper. At this point, a quantitative analysis needs to be conducted which is the last and a continuous phase.

4.2.2 Metrics

There are 3 main metrics that are used to evaluate the success of on-premise DLP improvement.

The number of the false positives (FP/TP)

It is very important that the detections are accurate and correct. False positive means that the scanner will detect a file as containing sensitive data, but it does not really contain sensitive data. If a scanner has a lot of false positives in the detections, it causes troubles in producing a lot of abundant detections that can consume time, energy, money and employee resources. Therefore, it is important that the number of the FP is as low as possible.

The number of the false negatives (FN/TN)

It is also important to make certain that all the files are inspected and all the files that contain sensitive data are identified. If a file contains sensitive data but does not become identified, it is called false negative (FN). The number of the FN is important to estimate how many files containing sensitive data may have been left undetected. The lower the number of the FN, the less sensitive data containing documents remain undetected, the more coverage of sensitive data in the discovery, and the more coverage in the protection. Consequently, by protecting the sensitive data containing documents that are detected, DLP will be improved.

The number of the alerts

Alerts are the notifications for the security analysts to inform them that a file containing sensitive data has been found. In a large organization, there are a lot of activities by the end-users on different files. If there are alerts for each of those files, it will be overwhelming for the security analyst to apply protection on all those files. This is the reason why it is important to set the alerts for more important files so that the security analyst can prioritize applying protection on the documents that are more important. This does not mean that the other documents are not needed to be protected. However, the files that contain a greater number of sensitive data are in priority.

5 Chapter 5: DLP Organizational Process

Depending on the size of the organization, the actions that make a change in the organization, especially when there is a change felt by the end-user employees, require several considerations. Other than planning how specifically an action should be taken, the process of all the actions should prior be planned and considered. The larger the company is, the more complex the process is.

In this chapter the whole process taken throughout this study will be introduced and explained as well as the lessons learnt and the limitations that affected the process significantly.

Microsoft introduces a process phases of Discover, Classify, Protect, and Monitor which inspired the author of this study [27].

The author defined each step for each environment in a way that some of the steps could be taken in parallel for both environments. Some of the actions depend on other actions to be taken, some actions can be done independently, and some actions can be taken at any time of the process. The action research phases taken inspired by the vendor are in figure 9.

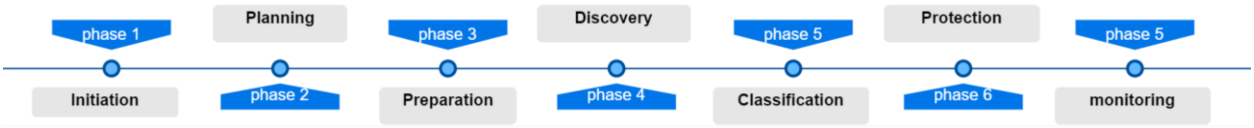


Figure 9. The process of Data Loss Prevention on both on-premise and endpoint environments.

Each phase includes some tasks that will be defined in terms of actions, results, corrective actions, parties involved, problems and lessons learned.

5.1 Initiation (on-premise and endpoint in parallel)

The initial step of the action research starts with research regarding both on-premise and endpoints. This research includes considering the whole picture from different organizational perspectives.

Before finding a solution, the problem to be solved needs to be identified. In case the problem related to any size of the company is lack of DLP, data discovery tool, inappropriate DLP path, or unidentified sensitive data across the company, this paper's process can be of use.

Learning the product (research)

This research may seem simple, but it is one of the most important considerations that if done wrongly, it will waste human, time, and financial resources. It includes readings, researching, analyzing and predicting which requires spending time and effort. The most common reference for researching about the product is the vendor's knowledgebase. For analyzing the product, the vendor websites may not be a complete reference. No vendor tends to disadvertise or pose against their own product. The disadvantages of the product should also be considered. There are different methods to analyze the products. Other than researching through search engines, testing them in a dedicated environment and Attending vendor seminars are two ways to identify the capabilities as well as the disadvantages of a product. Learning the product includes five sub tasks as below:

- Solution, feasibility, organizational requirements: After the identification of the organizational problems and the benefits that a selected/already purchased product can bring to the organizations, a solution is required to be decided. This thesis's author came up with some solutions as hypothesis. This hypothesis then was utilized as the basis of a questionnaire. For example, one of the solutions is configuring automatic or recommending a label to end users to prevent forgetting to label a document. There are other solutions related to endpoint DLP that are discussed already in Chapter 4 methodology. For on-premise lack of a discovery tool, Data Discovery tool named AIP scanner was used in this thesis as a solution.

  Feasibility includes taking into consideration how easy/difficult the implementation of a product can be depending on the context. Benefits of the product in the context of the organization is also important to be identified.
- Product challenges: All the possible challenges from organizational perspective, product related perspective, managerial perspective, as well as technical perspectives must be identified as much as possible at the initial phase. Nonetheless, some challenges will be identified while implementing and they may differ case-by-case.
- Proof of Concept AIP/SCC: In the case of this paper, the product necessary license for AIP and SCC had already been purchased and only needed to be implemented. However, In case a security specialist needs to decide about what product to purchase, there must be a research

through the market. The necessities of the company, and the capabilities of different products must be analyzed and the best product that meets the needs of the company must be identified. This selected product and the analysis must be presented to the related managers to prove the accuracy for the company and gain management support for the purchase of the product.

Even though the license in the subject company was already purchased, there was a need to understand which portal of AIP or SCC should be used.

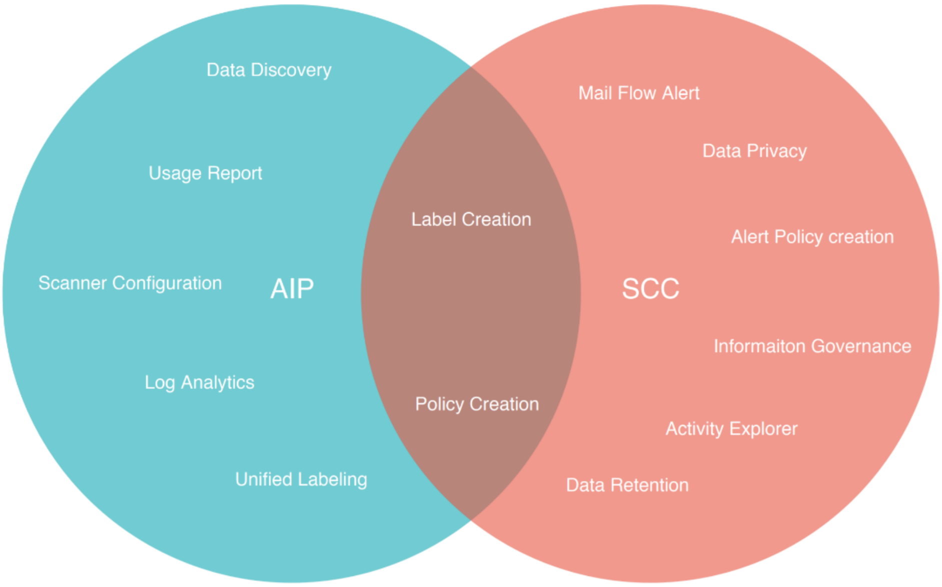


Figure 10. AIP and SCC features.

The author researched through vendor’s documentation and concluded that both portals complete each other. Figure 10 shows the features they have in common and the additional features they provide individually.

- Information Collection: Information about how the product should be purchased, how much does it cost (this mostly is performed by project managers in large organizations), how it should be implemented technically and so on, will fall into this category. The security specialists need to be well informed about the product as if they have produced it. Information about system requirements and architecture will also be collected in this sub task.
- Documenting the findings: Research findings needs to be recorded and documented for later reference. It can be in the shape of an internal organizational knowledgebase for introducing the project or can be a research diary.

- Planning: According to the findings and organizational status, the author planned the whole phases and tasks. Planning is divided into 2 sub plans for the two environments of endpoint and on-premise.
- External kick-off (action): As a small part of the process, after gathering adequate information from different perspectives, a presentation may introduce the whole process externally to other relevant departments or persons. In the case of this study, the Security Operation Center (SOC) manager, DPO, Head of Security and IT managers of each geographical site were invited to a meeting and were presented about the project. It is both providing brainstorm in terms of ideas and questions as well as defining the phases to clarify when in the project, there is a need of cooperation of whom. In large organizations, when a great deal of change is planned to be applied and there is a need of participation of different parties, a kick-off presentation with the participants included in the meeting needs to be conducted to agree upon timelines and clarify the project.

5.2 Planning

5.2.1 Endpoints

The labels as well as the policies, the configurations and the time of those configurations should be planned at this point based on the replies related to the questionnaire that was explained in Chapter 4 thoroughly as well as management decisions.

In addition, there must be some initial tests done to identify the advantages and disadvantages of the product which is chosen to be utilized. The tests and the analysis will be explained in more detail in chapters 6.

The author conducted tests to both practice and become a master in the product. This is the last phase to decide if the product functions appropriately. There were total of 8 tests that were identified to be of use to the company in the initial researches. 3 team members including the author were selected for initial tests. Member A owned a MacOS device with built in agent, member B and the author owned a windows 10 device with manually installed agent. The tests in detail will be explained in chapter 6.

5.2.2 On-premise

System requirements and architecture that was identified in the last phase will be applied and planned in the context of the subject organization. For example, the number of the targeted geographical sites, for scanning for sensitive data and planning to start with a selected site will happen at this stage. The scanner’s host windows server system requirements such as Random Access Memory (RAM), Hard Disk Drive (HDD), and virtual Central Processor Unit (vCPU) can be gathered from the vendor knowledgebase as well as database requirements. The connections, IP addresses and ports used can also be identified in the vendor’s knowledgebase.

5.3 Preparations

After gathering information and testing the product to be used, it is time to prepare the prerequisites. Preparations also fall into 2 environment (endpoint and on-premise) specific preparations:

5.3.1 Endpoints

All the labels’ and policies’ configurations that are suggested by the author were recorded in an online document and shared with relevant parties such as DPO and the head of security to have their comments and take corrective actions accordingly. Labels including automatic label was suggested first which was not approved. The configurations were complete meaning that all the configurations of user restriction, encryption, and automatic labeling were all suggested at once, which was then found not appropriate by the author after not receiving the approval.

The main important fact is that a sudden big change that creates a shock on the employees is not appropriate. The change must be applied gradually. For instance, first the labels become created simply without protection or automatic labeling and become tested on a small group, then gradually extend and expand to other changes as well as training the users when needed. Therefore, the author designed another label taxonomy accordingly.

5.3.2 On-premise

Local IT department was contacted through the company’s ticketing system to install the windows server. This is recommended to be done very soon since different specialists should take each part and configure according to their defined responsibilities. For instance, a VM must be created by the Vcenter owner, the user by which the scan authenticates itself to AIP, must be created by the system administrators, and the connections must be opened as a firewall rule by a security member who is responsible for it. Therefore, such types of tickets that should be passed through several system owners take a long time to be case closed. The Database stores the configurations while installing the scanner and when the scanner becomes installed, a node becomes automatically created in AIP portal by the installation script.

5.4 Discovery

5.4.1 Endpoints

As soon as the agent is installed on the windows 10 endpoints, the labels will be downloaded and applied on the office products as a plug-in on endpoint machines (MacOS and Windows 10) based on the configurations of the labels and policies.

5.4.2 On-premise

When installation is complete, it is time to initiate the first scan. It is very important to record all the findings at this point because all these actions will be used in the future for the other major sites' scanners. The best method to troubleshoot problems when starting a scan is to check the logs. It is either the product that provide connection logs when conducting a scan, or some standalone tools such as Wireshark can be used to troubleshoot connections. The vendor's customer support can also be an option.

After the installation of the AIP Scanner host server, the author initiated fine tuning the configurations on the AIP portal. There are some actions taken while reading the instruction provided by the vendor such as creating a content scan job. After the installation of the scanner, the author used the commands that were provided from the vendor to initiate the very first discovery scan.

The scan was scheduled for an appropriate time with less network traffic of data to prevent interruptions with other services or systems running through the network such as the weekends or after working hour. IT managers were informed about the scans before the scan. To validate if the scan was successful, scan logs as well as the logs belonging to the files that are scanned can be checked and analyzed.

The author Decided to initiate with a small scope of scanning a single file server with 31 shared repositories on one geographical site which will be covered in this thesis. The reason to start with a small scope is to explore any obstacles and challenges and do ability of the scanner preparations, installation, and conducting the scans. This method provides more confidence in the future deployments across all the other sites. Other major geographical sites are out of the scope of this thesis. At this point the author only conducted discovery scans.

5.5 Classification and labeling (both on-premise and endpoint environments)

After the discovery of all the files in both repositories and the endpoints, using any product, the logs should be visible by some means. In the case of this thesis, as soon as the repositories are scanned, all the files discovered will be logged in the activity log in AIP.

The files that were already labeled, the files that were detected as containing sensitive data, and the files without any label applied will all be gathered in Activity log which has the capability to filter the files based on the label and sensitive data type detected. Data Discovery and classification need to be continued until all the 5000 endpoints and all the file servers are covered.

5.6 Applying protection (on-premise and endpoints same labels)

It is possible to apply protection configuration while creating the labels and the policies. In the case of first time creating the labels to be published for the endpoints, discovery and protection can be implemented at the same time. It does not make any difference in terms of the necessity to train the end users before publishing the labels in either case of brand-new labels or improved ones. In both cases, the end user should become informed and trained beforehand.

The same labels are used for both on-premise and endpoints. However, the difference is that even if the protection is configured in the labels, they will be applied on the documents on the endpoints but for the file servers, the scanner needs to conduct both discovery and protection scans to be able to apply protection. Therefore, even though the labels are configured to protect documents, as long as the scanner scans to discover only, the protection will not be applied on the files in the repository.

There are several choices at this point:

- Automatically applying and protecting a label on sensitive data containing documents.
- Applying a default label on all the documents. A neutral label that is neither too restrictive and protective, nor without any protection. The author decided that “Internal” is an appropriate label as a default label.
- Identifying old files, classify, and apply protection/retention on them.
- Exchanging the files that were already applied by the old labels with new corresponding labels. For instance, a newly created label can be replaced with a similar old label on the documents.

6 Chapter 6: Analysis and Result

6.1 Initiation

6.1.1 Endpoints

Out of 649 employees, 192 employees responded the questionnaire in the first round of the announcement. After the reminder email, the number increased to 221 employees. Therefore, reminder email was affective in increasing the number of responses.

The questions started branching halfway through. Therefore, the respondents to the questions were filtered to lower numbers moving forward in the questionnaire. Branching means that for example, if questions 1, 2, and 3 come each after another, depending on the answer that the respondent chooses for question 1, the questionnaire directs the respondent to question 2 or skips question 2 and directs the respondent to question 3 as the next question. (See Appendix 1)

Out of the total respondents, (Figure 11)

1. 30 (13%) employees have been working in the subject organization for less than a year.
2. 61 (27%) employees have been working in the subject organization from 1-3 years.
3. 82 (37%) employees have been working in the subject organization from 3-10 years.
4. 48 (21%) employees have been working in the subject organization for more than 10 years.

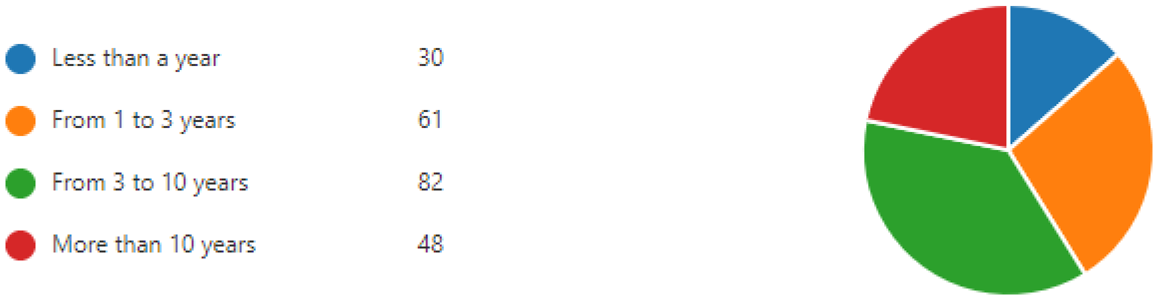


Figure 11. The length of the time that respondents have worked in the subject organization.

The questionnaire was designed for the purpose to determine the find 10 categories of information which are explained in detail below:

Identifying The departments that are directly in contact with sensitive data.

Only 3 (1.3%) people out of 221 respondents have replied to be dealing with sensitive data in 100% of their documents. These three employees where in 2 different departments. One of these three departments is HR and the other two were from a department called as the name of a product of the company and for confidentiality reasons, it is named differently in this thesis. The author names this department as ProductOneDepartment. ProductOneDepartment is a production department that deals with producing the main product of the company that is mostly dealing with customers related to Business to Customers (B2C). B2C is directly in contact with individual customers' data.

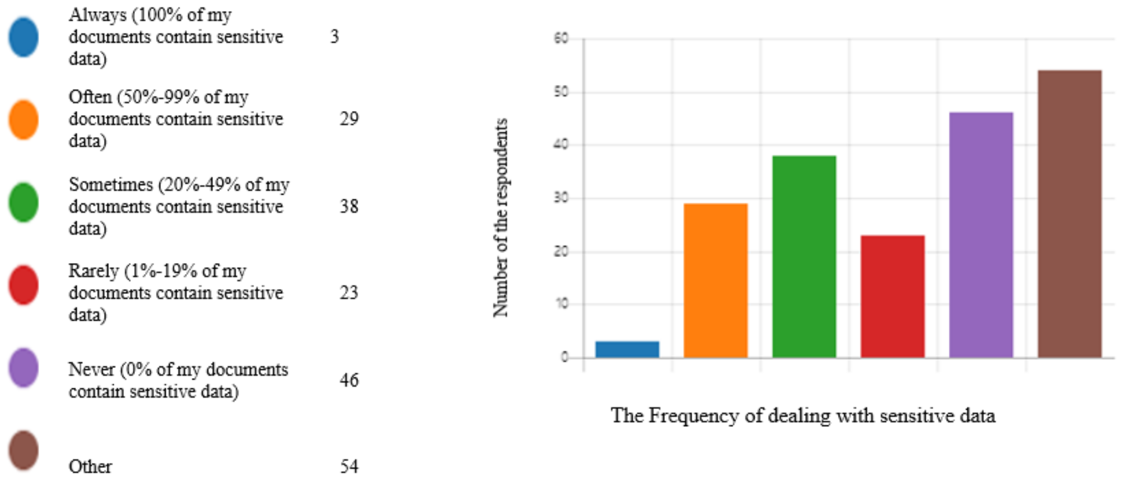


Figure 12. The frequency of the respondents dealing with sensitive data.

The results showed that regardless of the frequency of dealing with sensitive data, the highest number of the respondents who dealt with sensitive data from 1% to 100% came from departments of HR (4 responds) and Service operations (7 responds) which interact with employees and customers Business to Business (B2B) respectively. Service Operation and HR are different from the aforementioned ProductOneDepartment (B2C).

24 different respondents with different position titles claimed that they often deal with sensitive data (50%-99%). These 24 employees were in 11 different departments. The author did not expect this varied output from only two geographical sites out of 8 major sites. The expectations were at the very most 3 departments to be often dealing with sensitive data. This is because when the author had talked to different managers, the managers had advised on 3 main

departments that were supposed to deal with sensitive data. HR and Service operations and ProductOneDepartment were predicted and proved true in the questionnaire's result in highly dealing with sensitive data. Thus, this variety of 11 findings should also be predicted to happen on other geographical sites which will bring more complexity in larger scope.

Lessons learnt: One of the most valuable pieces of information in the result was that the author was mostly searching for the departments that deal with sensitive data initially. However, after the questionnaire responds were collected, the author found that in one department, some employees deal with sensitive data while some do not. This means that dealing with sensitive data is not only related to the department in which the employee works, but also the position title of the employee can represent their connection with sensitive data.

Since there were varied number of the departments and positions titles that deal with sensitive data, management decided not to use department special labels. This is because in case each department has its own label, the complexity from the backend administrative configuration would cause confusion. For example, a label that is created for HR may appear in another department because of misconfigurations and the complexity of planning for all the departments. It will also cause confusions from the end-users' side in terms of which label to choose for their documents.

Out of 221 respondents, 147 respondents claimed to be dealing with sensitive data (always, often, sometimes, and rarely). Out of 147 respondents who deal with sensitive data, 18 respondents either did not mention their departments, or left the entry with * character, or their department was mentioned less than 3 times by different respondents which are illustrated as unclassified in Figure 13. Figure 13 illustrates the numbers of the respondents who deal with sensitive data, regardless of how often they deal with sensitive data, in different departments [45].

The number of the employees who are aware of protection labels out of total respondents.

In total, only 28 out of 221 respondents were not aware of the protection labels 5, 6, 11, and 6 of whom have been working in the subject organization for less than a year, 1-3 years, 3-10

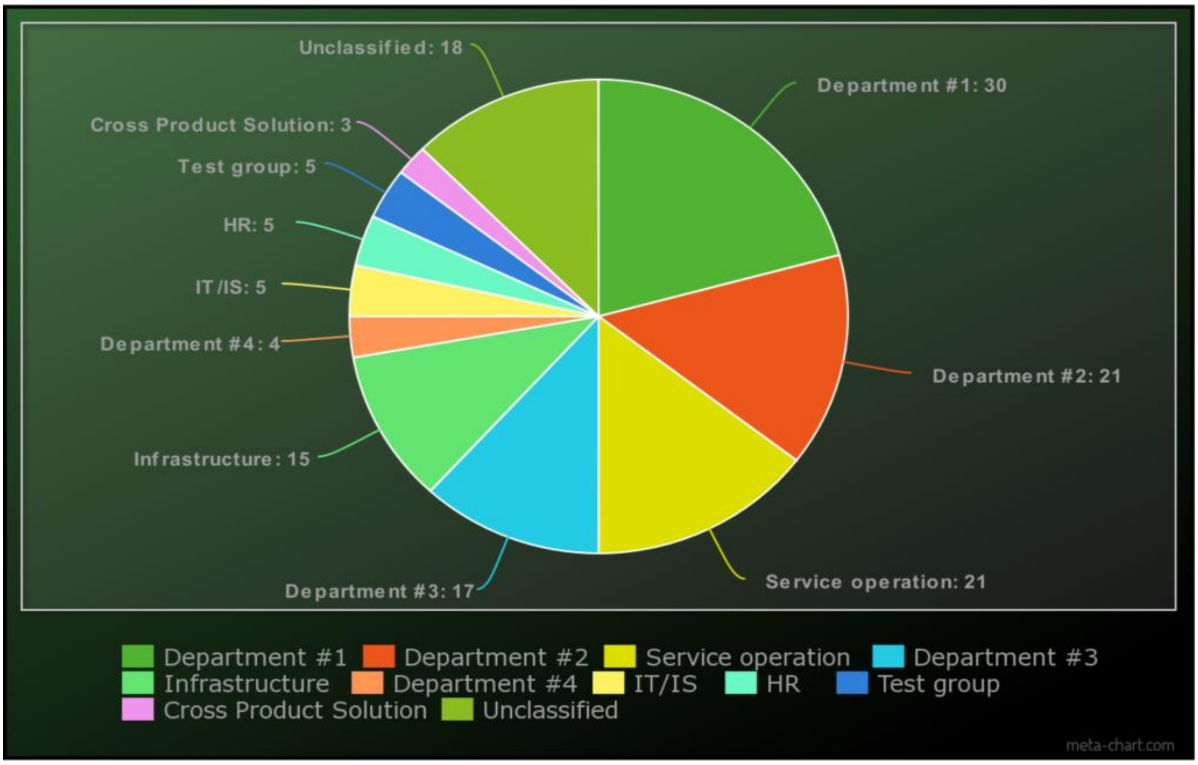


Figure 13. The numbers of the respondents who deal with sensitive data, regardless of how often they deal with sensitive data, in different departments (Online tool used [43]).

years, and more than 10 years respectively. This result means that, knowing the labels does not depend on how many years the employees have been working in the subject organization. It either reflects the lack of existence in training all the staff by either the company academy department, or by their direct managers.

The number of the employees who use the protection labels on sensitive data containing documents.

32 respondents deal with sensitive data more than 50% of the times. Out of those 32 respondents, 4 employees never use the protection labels on their documents. One out of these four employees claimed that the lack of training is the reason for not using protection labels. Other 3, selected “other” as other reasons why they do not use the protection labels.

All in all, out of 221 respondents, 147 respondents claimed that they deal with sensitive data more than 0% meaning that they did not reply that they never deal with sensitive data. Out of 147 respondents, 91 respondents fall into the employees who deal with sensitive data and not always used protection labels for those of sensitive data containing documents. As illustrated in Figure 14, out of those 91 respondents who deal with sensitive data and not always used protection labels, 26, 21, 12, 10, 9, 6, 3, and 2 respondents selected forgetting, no mind, other reasons, confusion, lack of training, weak protection, Availability limitation, and time waste respectively for the reason why they may not use the protection labels.

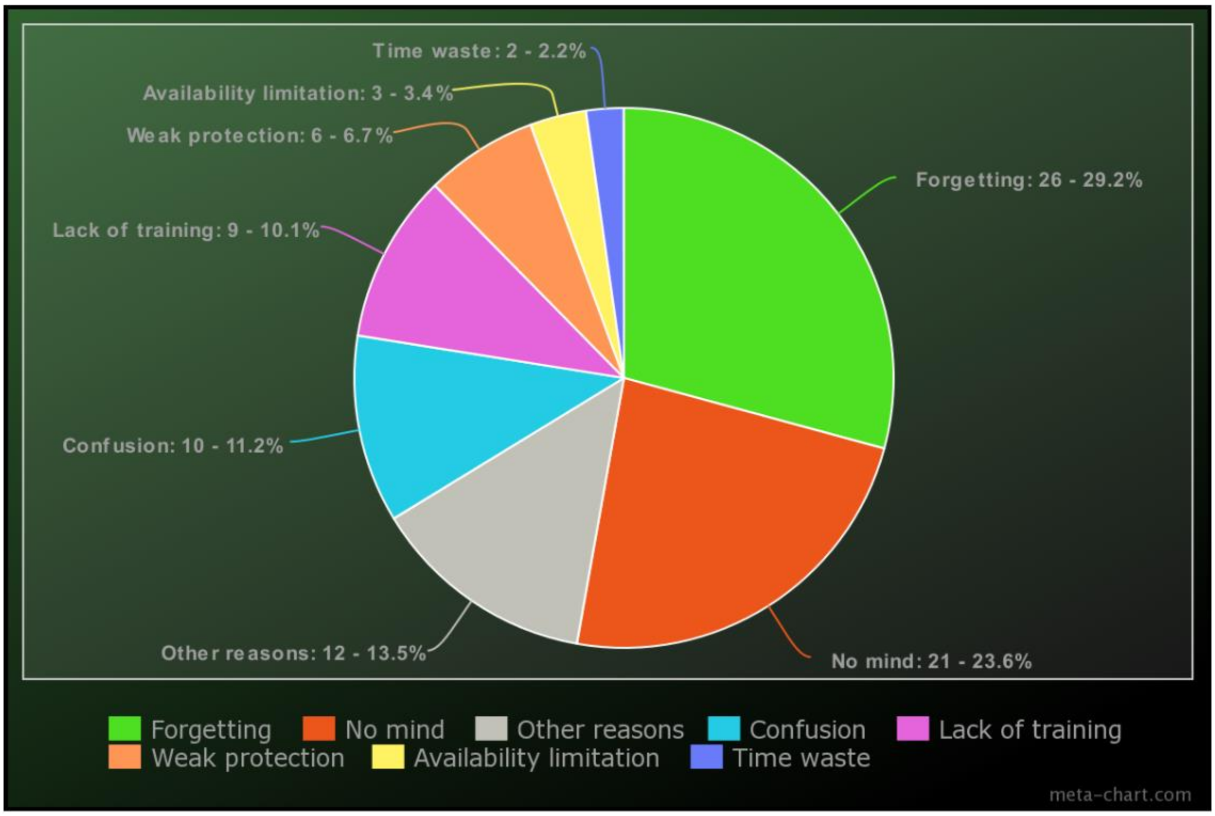
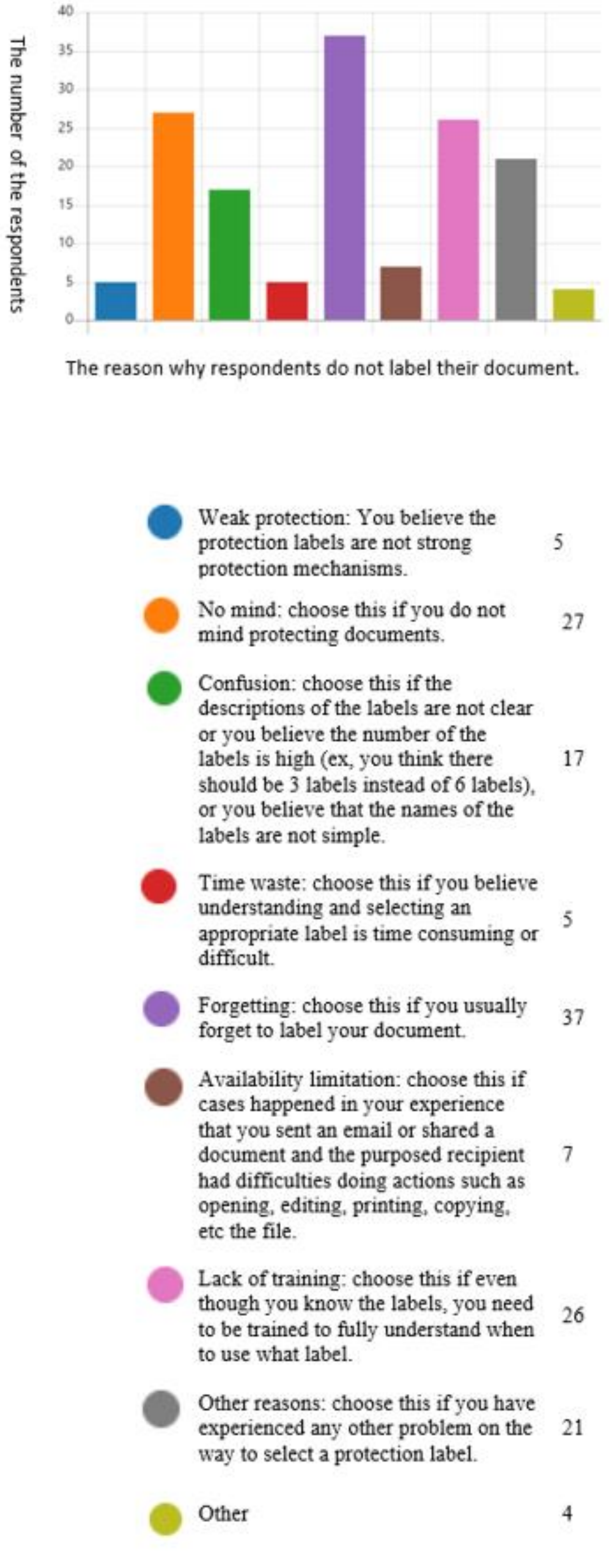


Figure 14. The percentage of the reasons why the respondents do not use labels [45].

Out of those 147 employees, 17 employees never use the protection labels. Out of them, 1, 2, 3, 3, 5, 1, and 2 people selected confusion, forgetting, lack of training, no mind, time waste, other reasons and no reply respectively.

Out of 106 respondents who had reached the branch of question 8 which allows more than 1 option, 37 respondents claimed forgetting to label documents is the main reason why they may not label their documents (Figure 14).

The availability limitation extension of protection labels.



●	Weak protection: You believe the protection labels are not strong protection mechanisms.	5
●	No mind: choose this if you do not mind protecting documents.	27
●	Confusion: choose this if the descriptions of the labels are not clear or you believe the number of the labels is high (ex, you think there should be 3 labels instead of 6 labels), or you believe that the names of the labels are not simple.	17
●	Time waste: choose this if you believe understanding and selecting an appropriate label is time consuming or difficult.	5
●	Forgetting: choose this if you usually forget to label your document.	37
●	Availability limitation: choose this if cases happened in your experience that you sent an email or shared a document and the purposed recipient had difficulties doing actions such as opening, editing, printing, copying, etc the file.	7
●	Lack of training: choose this if even though you know the labels, you need to be trained to fully understand when to use what label.	26
●	Other reasons: choose this if you have experienced any other problem on the way to select a protection label.	21
●	Other	4

Figure 15. The reason why the respondents tend not to use the protection labels.

Only 3 respondents out of 221 respondents claimed that the existed labels limit availability from the receivers of the documents. (See Figure 14 and 15)

The purpose to measure availability limitation was to identify the problems that may occur while using the already existing protection labels. This was included as one of the assumptions. The result for this measurement rejected the assumption as there were only 3 respondents out of 221 respondents claimed that the existed labels limit availability from the receivers of the documents.

Deciding if the employees need training for using the labels

In total, 91 out of 221 respondents required training which is more than 41%. 46, 38, and 34 respondents required training to learn 1. How to choose the best label for their documents, 2. How to use and choose the best labels, 3. Where to find, how to use and how to choose the best label for their document respectively (Figure 16). These options are not exclusive, but they were asked in order for the author to understand what should be included and more focused in the content of the training to be conducted. The author decided to create a document and place it in the general knowledgebase platform of the organization, announce and make the document available for the employees.

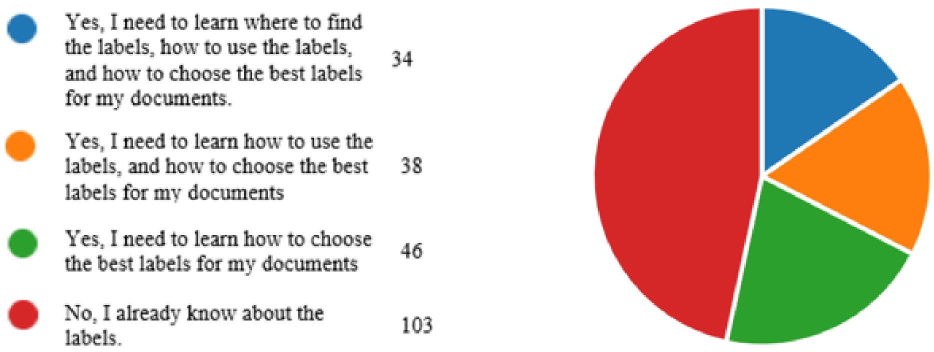


Figure 16. Training needs.

The number of the employees who are aware of internal related policies

10, 18, 150, and 43 out of 221 respondents mentioned that the internal organizational policy related to classification of documents was unknown to them, they have heard such document exists but they did not find time to read it, partially, and know it very well respectively. (Figure 17)

This result means that more than 80 percent of the respondents either have not read or made aware by any means about the internal policy related to data classification. The decisions based on the results of the questionnaire can be an annual enforcement to read the related internal policy in order to meet 2 purposes:

1. The employees who have been working for more than a year in the subject organization become reminded of the responsibility of protecting information,
2. The employees who have been working for less than a year in the subject organization could become familiar with the protection labels. This can also be met in case a detailed practical guide or internal policy related to protecting documents become created and obligating the new coming employees to read and learn it by heart.

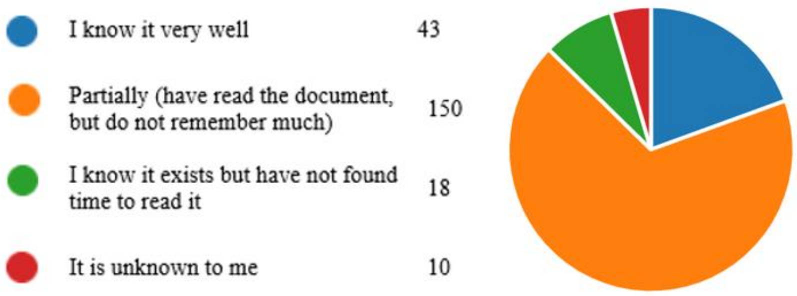


Figure 17. The number of the employees who are aware of internal related policies.

The complains/concerns of the employees about the current status of the labels

The author herself experienced forgetting to label very important documents for several times which was the motivation to propose if forgetting to label a document might be one of the reasons that prevents employees to use the protection labels.

Even though forgetting to label the documents broke the record of 37 responses, there were also other concerns from the end-user's side.

The bar chart in Figure 15 illustrates the other reasons. Not minding assigning a label stands in the second highest selected option which means that firstly, the respondents were honest, and secondly, they simply do not care labeling their documents.

Forgetting to label at the very least documents that contain sensitive data will be solved by recommendations mentioned in purpose 11. lack of training in which label to choose will be solved by annual information security training. Confusion will be resolved by simplifying the label taxonomy.

The employees who believe that it is a waste of time or they do not mind labeling their documents may stay unresolved and may be remained for further study.

The suggestions of employees

Even though the final decisions will be made by the management; it was also interesting to hear the voice of the end-users to rise ideas in better tuning the labeling taxonomy. The respondents were asked to give their opinion if they prefer to have their own department-specific-labels or not. 142 respondents who had passed through some branches, replied in total to this question out of which 57% disagreed and 43% agreed to the idea.

As of a large organization, the design of specific labels for each department becomes very complicated due to the fact that there are varied departments with different position titles that different frequency of dealing with sensitive data. To test if it is beneficial to create department specific labels, there is a need to test at least one employee out of each position title and department to test if a specific label for that department is recognized appropriate or not. This was not possible and was similar to real deployment rather than testing since there could be so many employees involved. In addition, it required a lot of effort and time without assurance that it eventually will be approved by the management or not. Therefore, the author decided to take the respondents' suggestion and not to attempt designing, testing, and deploying department specific labels.

Deciding if it is better to have a new set of label taxonomy

Only 16% of the total 106 respondents who passed through the branches, were confused to choose an appropriate label for their documents.

Based on the result of the questionnaire and the managerial opinion as well as the experiences gained by the author as the technical lead of the implementation, 5 labels were created by the author to the management including 3 of the former labels (Public, Internal, and confidential) , as well as 2 other labels named Non-business and Sensitive. All the labels other than Sensitive were set to be manually applied by the end-users without prompting the user or any obligation in applying the labels, and a separate label specifically for sensitive data containing documents which was designed to detect sensitive data based on the pre-defined templates by AIP tool and only warns the users about the content and recommends the user to apply the label. The user will either accept the label to be applied on her/his document or ignores the prompt and justifies the ignorance. The logs of justification will be collected by system log and sent to Azure Activity log for administrative visibility. Force pushing automatic labeling was not approved or applied. The new labels were Non-business, Public, Internal, Confidential, Sensitive (more details of the label setting plan can be found in Appendix 5):

Determining if there is a need to have default label, automatic labeling assignment, label reminder prompting.

As mentioned earlier, forgetting to choose the labels was the highest selected reason why a user might not label their documents. This result means that the users need a solution to be reminded of labeling sensitive data containing documents. One solution is to prompt the users by some pop-up to warn them that their document contains sensitive data, and another solution is to force push automatically label the sensitive data containing documents and only inform the users that because of their document containing sensitive data, their document has been automatically labeled and protected. Setting a default label which is applied on all the documents is another way to apply labels on all the documents. Up to this point of time, setting a default label is only proposed by the author but has not been approved by managers in the subject organization.

6.1.2 On-premise and endpoints

The initial problems related to the subject company can be researched by contacting different employees such as GRC members, IT, DPO, head of security and direct manager. The author contacted all the mentioned parties and identified the overall organizational problems to be solved.

The most important problem related to on-premise file servers were that there is no data discovery tool in the company to determine where sensitive data resides. The problems related to both on-premise and endpoints are as below.

1. Inconsistencies:
- Labels in both cloud services
- Between the label bar and sensitivity description
2. Lack of standard label namings
3. User difficulties
- Users may forget to label documents manually.
- Lack of default label
4. Lack of Data discovery tool

1. Inconsistencies

For every company the current status of DLP must be analyzed and it is inevitable that there cannot be a single status for all the companies. Nonetheless, the status of different organizations’ DLP may fall into 2 Categories:

a) The organization has already had an attempt in implementing DLP. In this situation, the security/IT specialist is required to improve the status of DLP.
b) The organization is just initiating DLP implementation, and there has not been any attempt into labeling and protecting the documents. In this situation, the security/IT specialist must kick start the whole process from the beginning. In the context of this research, this situation is less complicated that the first status. This is due to the fact that there is no need for additional research through what has been done so far, what should be corrected, or what should be continued from the previous attempt.

This thesis falls in category (a). After analyzing the current status, inconsistencies were found. As explained earlier, the both management portals of SCC and AIP have the feature to configure labels.

Labels in both cloud services (AIP and SCC portals): In the subject organization, IT department had configured different sets of labels in each portal which had caused challenges. There are other features in each portal that inherits these configurations. For example, AIP has the capability of discovering the documents and labeling them automatically. Logs of the user activities will also be visible in AIP portal. SCC is mostly providing features for Data retention, email flow, information Governance, DLP Alerts, etc.

All these features, if the organization intends to leverage them, should be consistent in terms of the basis that they are working accordingly. In other words, all these features in both portals need to function based on a single labeling taxonomy.

Due to this problem, the IT specialist had to create any label that was created in AIP, in SCC manually because these two portals did not sync with one another automatically.

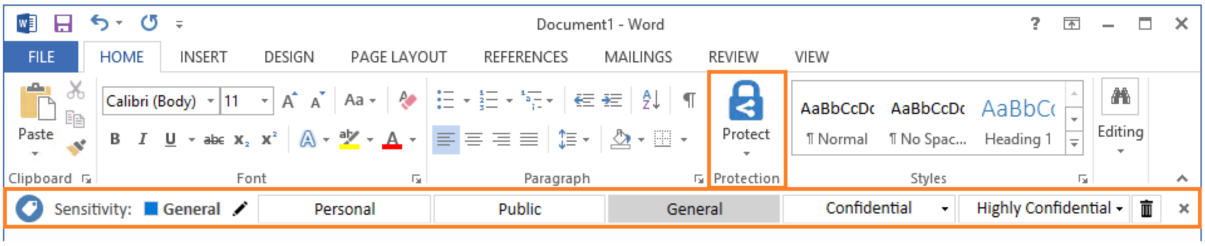
Between the label bar and sensitivity description: On the other side, from the end-user workstation, the labels that were on AIP were visible. This was becuase the client that was installed on the endpoint workstations, named AIP classic client, which downloads the labels from AIP and not SCC. One misconfiguration was found in AIP management portal that had led to an inconsistency between the general description of sensitivity label, and the labels that were revealed to the end-users. The description of the sensitivity label explains that there are four labels of Public, Internal, Confidential, and secret, while the labels published to the end users included 6 labels of Public, Internal, Confidential and 3 other labels. No label named

Secret had been published to the end-users. This Provided confusions for the end-users with two different descriptions.

2. Lack of standard label namings

As mentioned earlier, 6 labels already existed when the author initiated the DLP improvement. 3 of the existed labels were Public, Internal, and confidential which are rather standard. Other labels had the word encryption in the title name of the labels and the name of the subject company was also in the title names. If the name of the company is for example Velvet, The other labels were as: Velvet and public encryption, Velvet and confidential and read only, and Velvet confidential and encryption.

Even though each label had description, they are not considered to be standard. The names of the labels must be self-explanatory, clear, and short (at most 2 words). An example of a standard label taxonomy is in Figure 18.



Sensitivity: General

Figure 18. A standard set of label taxonomy [46].

3. User difficulties

- Users may forget to label documents manually.
- Lack of default label: The author proposes that if there is a label applied on all the documets that are created by the employees automatically by some administrator role (the author herself), DLP will be improved. On the other hand, if all the documents including the sensitive data containing documents become labeled as internal, it is not enough protection for them.

4. Lack of Data discovery tool- Lack of understanding what sensitive data is and where it resides.

Last but the most important problem that was identified in the subject company and is most probable to be identified in other cases, is the fact that there is a lack of a tool that could discover all the documents of the company, lack of understanding what sensitive data is and where it resides.

6.2 Planning

6.2.1 Endpoints

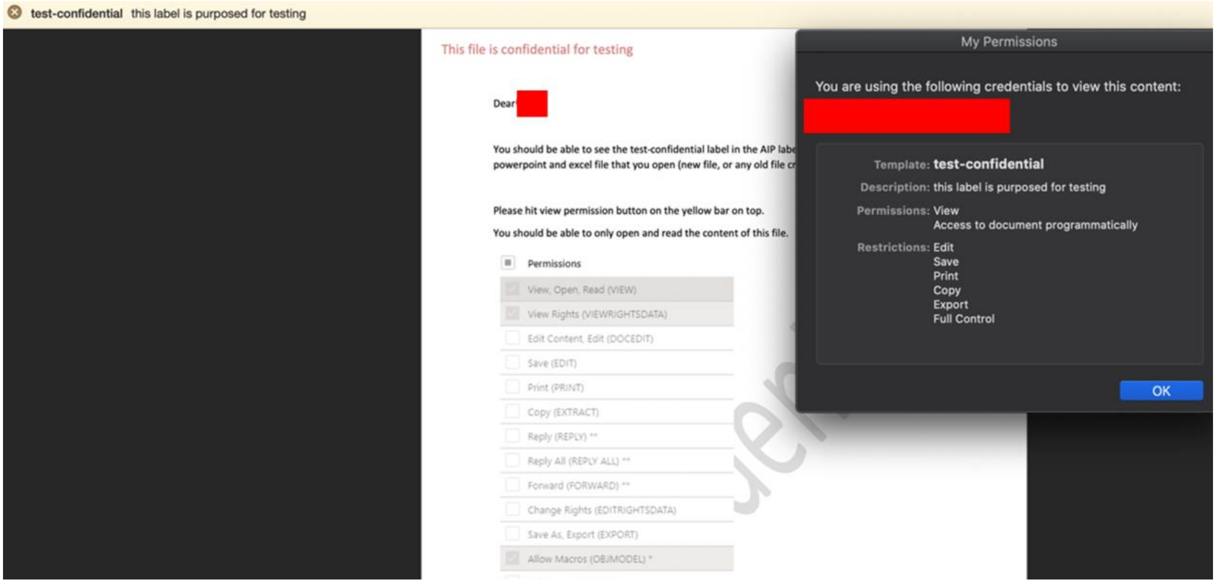
Test 1. Basic checks

Objective: A) Determining which portal is the dominant portal and understanding which portal loads the labels to the endpoints currently. B) Testing if the permissions work properly on both MacOS and Windows devices

In label configurations there is a permission section that defines the user permissions and this permission sticks to the label. When the label is applied on a document, the permission configured on the label will also become applied on the document. For example, if a user is given a viewer permission in a label named confidential, the user can only open and read the documents that are labeled as confidential.

2 simple labels with the same permissions and configurations, one in SCC and another in AIP portals, but with different names were created to see which name appears for the end-users in their office applications. Member A with MacOS was given a reviewer and member B with windows 10 machine was given a viewer permission in both labels. Figure 19 illustrates the user side’s view of the label that was named as test-confidential. As shown in Figure 19, the user who receives such document via email or sharing, will only have view permission.

At this point, the author figured that MacOS does not download the labels if the labels are created in AIP portal. Only if the label is created on SCC portal, and the unified labeling is activated, the label appears for MacOS. This is because MacOS has the built in Unified Labeling client which cooperates only with SCC. Whereas, the agent that was installed on all windows 10 endpoints already was the classic client (the author explained earlier that the classic client was already installed on all the endpoints way before DLP improvement process started by another employee in the past.) which cooperates with AIP. Even though the labels do not appear for MacOS machines, a document that is labeled in Windows 10 and shared with a user that has MacOS, carries and applies permissions for that document anyway. Permissions are configured based on user domains, and they do not depend on the host OS. Only the users who are using MacOS are unable to see the labels on their MS Office applications to use them on their documents.



test-confidential this label is purposed for testing

This file is confidential for testing

Dear [redacted]

You should be able to see the test-confidential label in the AIP labe
powerpoint and excel file that you open (new file, or any old file cr

Please hit view permission button on the yellow bar on top.

You should be able to only open and read the content of this file.

■ Permissions
☑ View, Open, Read (VIEW)
☑ View Rights (VIEWRIGHTSDATA)
☐ Edit Content, Edit (DOCEDIT)
☐ Save (EDIT)
☐ Print (PRINT)
☐ Copy (EXTRACT)
☐ Reply (REPLY) **
☐ Reply All (REPLY ALL) **
☐ Forward (FORWARD) **
☐ Change Rights (EDITRIGHTSDATA)
☐ Save As, Export (EXPORT)
☑ Allow Macros (OBJMODEL) *

My Permissions

You are using the following credentials to view this content:
[redacted]

Template: test-confidential
Description: this label is purposed for testing
Permissions: View
Access to document programmatically
Restrictions: Edit
Save
Print
Copy
Export
Full Control

Figure 19. basic checks.

Test 2. Auto labeling custom condition

Objective: Test if keyword detection works properly.

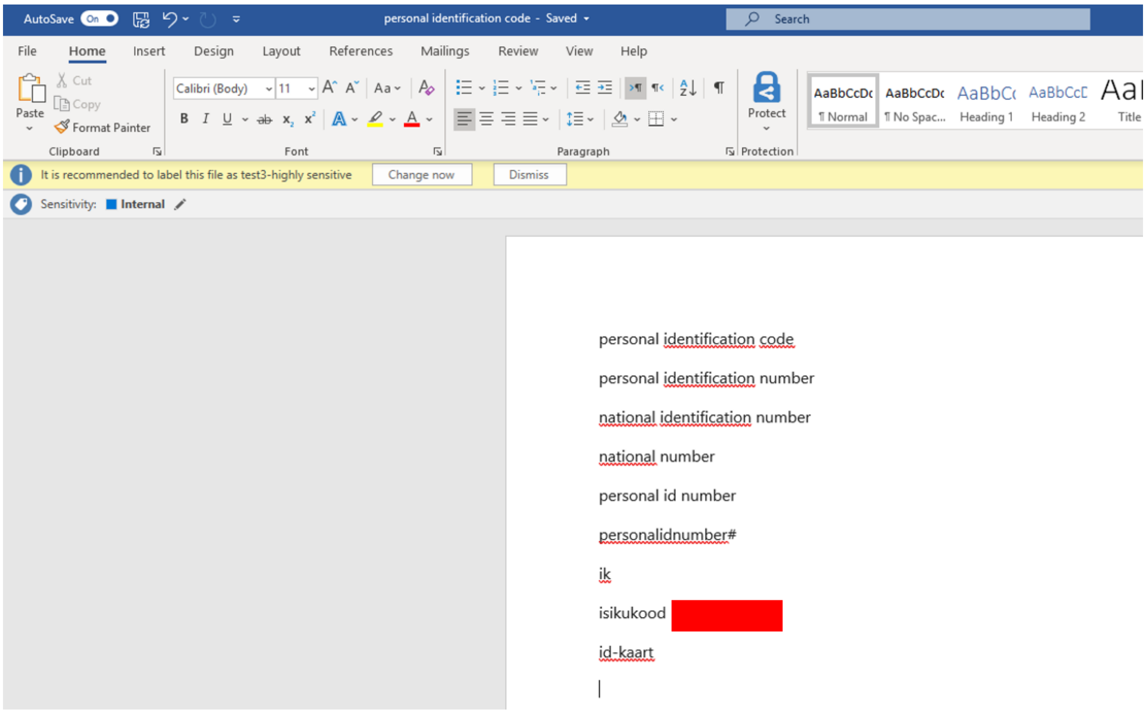
A simple keyword of "password" was given in the label configuration to detect on a document. The same permissions as in test 1 was given to the members. The author created the file with the content that had the word "password" inside. Right when intended to save the file, the author was prompted that the document was automatically given a label.

Test 3. Auto labeling template condition label recommendation

Objective: Testing if the predefined templates[1] to detect sensitive data works properly. In this test, the permissions were not set at all (all permissions were open), because they were not

[1] Predefined templates are built-in templates by design in the label setting which provides automatic detection of different sensitive data types. For example, there is a predefined template named Credit Card Number which can be set in the configuration of a label. This mechanism reduces the work for the system administrator to search and identify different types of sensitive data and create a RegEx or keyword for them and applying them in a label setting. Instead, Microsoft has done the research and provided the automatic detection of sensitive data types that are defined and regulated in different countries around the world and provided ready-to-use predefined templates.

purposed to be tested. The author intended to test if the predefined sensitive information related templates work properly when configured to only recommend the user to apply a label based on the content detected as sensitive contained.



It is recommended to label this file as test3-highly sensitive | Change now | Dismiss

Sensitivity: Internal

personal identification code

personal identification number

national identification number

national number

personal id number

personalidnumber#

ik

isikukood [redacted]

id-kaart

Figure 20. Auto labeling template condition label recommendation.

To address that, instead of giving a keyword, a predefined template specifically for EU debit card number was defined in the label configuration. The test file containing EU debit card number was detected correctly and Dismiss[1] was allowed for the author without justifying a reason.

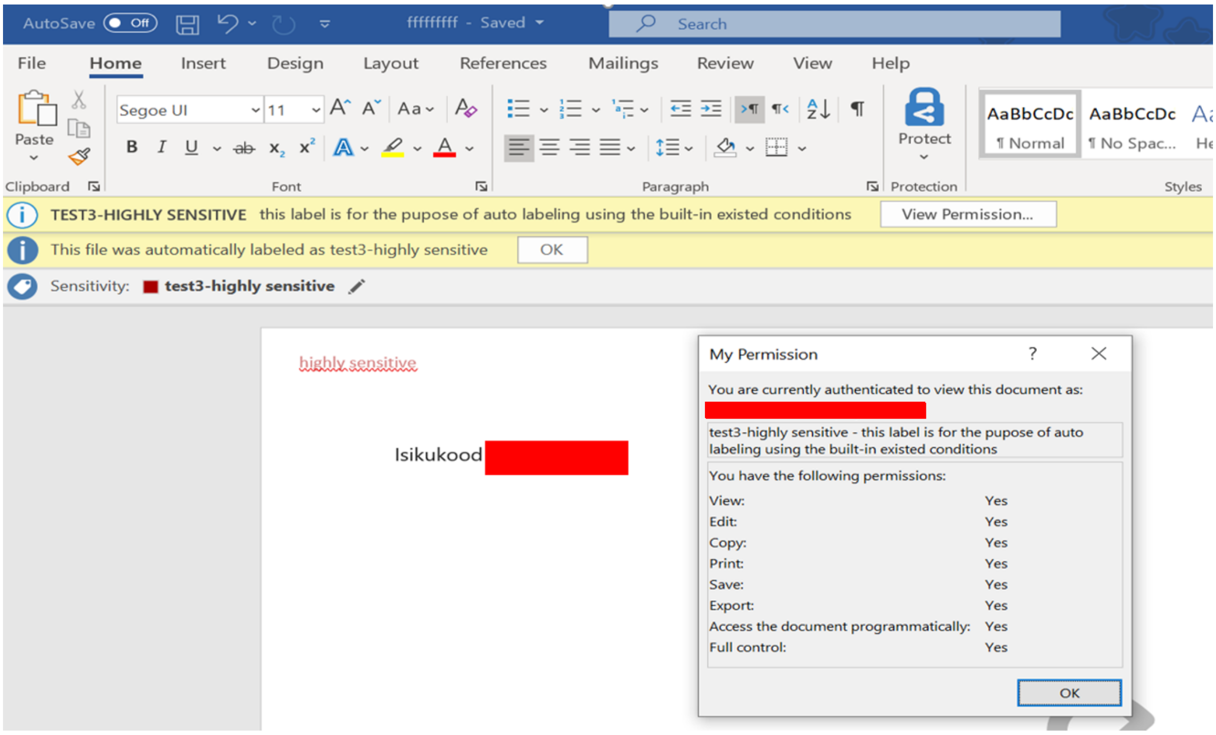
Test 4 Auto labeling template condition label enforcement

Objective: To test if the automatic labeling (forcing a label on a document that contains sensitive data) works properly.

This test was exactly like test 3 with the difference that the label applies forcefully. The user experiences an automatic labeling on the document and lowering the label to a less restrictive label needs justification. This justification was both visible in the event viewer on windows and

[1] Dismiss here means that the user has the option to accept the label that is recommended or ignores it with a button of Dismiss.

in Activity log in AIP portal. As illustrated in Figure 21, the label was automatically applied on a document which contained personal identification code which is red marked in Figure 21.



TEST3-HIGHLY SENSITIVE this label is for the pupose of auto labeling using the built-in existed conditions View Permission...

This file was automatically labeled as test3-highly sensitive OK

Sensitivity: 🟥 test3-highly sensitive

highly sensitive

Isikukood [redacted]

My Permission

You are currently authenticated to view this document as:

[redacted]

test3-highly sensitive - this label is for the pupose of auto labeling using the built-in existed conditions

You have the following permissions:

View:	Yes
Edit:	Yes
Copy:	Yes
Print:	Yes
Save:	Yes
Export:	Yes
Access the document programmatically:	Yes
Full control:	Yes

OK

Figure 21. Auto labeling template condition label enforcement.

Test 5 Label load to SharePoint online

Objective: To test if the label travels with the document when it is uploaded to SharePoint.

A document was created by the author and applied a label with the same permissions as in Test 1 for employee B. The author then uploaded the document on SharePoint and shared the document with employee B. The result was that employee B could only view the document online and could not save or open it on desktop.

Test 6 Label load to OneDrive for business

Lessons learnt: The same for Sharepoint applied and resulted for Onedrive. The difference between Onedrive/SharePoint and office applications of Excel, Word, PowerPoint and outlook is that the labels do not appear in Onedrive/SharePoint. However, if a file is created on desktop and uploaded on Onedrive/SharePoint, the label travels with the document and protects it. The problem was that if a document is created directly online from Onedrive/SharePoint, unfortunately they cannot be protected from there since the labels do not appear in Onedrive/SharePoint.

Test 7 organizational specific confidential data

Objective: Testing if RegEx works properly.

In large organizations, some information with the same format may be defined as confidential and require protection. In case the information can be defined by a RegEx, it can be configured in a label to detect and apply on documents. A RegEx was configured and worked properly in detection regarding a specific type of information in the target company. Due to confidentiality considerations, the test and the result is not illustrated in this thesis.

Lessons learned: Regex can be configured in AIP but there is no configuration of RegEx on SCC. Therefore, when the Unified Labeling client is installed due to deprecation of classic client on all the endpoints, there will be no feature of RegEx to leverage.

Test 8 sub-labeling

In this test, a very simple sub-label was created to see how it works. It appeared on the office applications as a drop-down by clicking on a parent label. The sub-label inherits the configuration of the parent label and it can have more configuration for itself. When a parent label has a sub-label, the parent label cannot be selected by the end-users. Instead, a drop-down menu with the sub-labels can be selected.

6.2.2 On-premise

The scanner was planned to be installed on top of a windows server 2019 host virtual machine. There are varied number of file servers in a large organization. However, determining which file server to choose for the first scan is important. Because based on the first scan, it will be decided if there is enough reason to continue the whole implementation on the other file servers.

The author decided to choose the least critical and the simplest situated file server for the first scan. Simplest situated means that if the parties on the site related to the file server are more cooperative and fast in providing the services needed. Critical means that in case of some mistake in the classification, or any risk or thread to the windows server host machine, or the account used for the scans, there would not be a lot of highly risked damage to the file server and the company as a whole consequently. Therefore, the author already knew that the file server selected for the first scan may not contain a lot of sensitive data containing documents. This first file server is usually chosen either by the direct manager of the security specialist, or as IT department is most familiar with the systems, they can be the focal point for giving the least critical file server FQDN or IP address.

Lessons learnt: In a large organization, there are varied number of file servers on each site. Some sites have only one fileserver and some more than one. In case there is only one fileserver on a site, one scanner can be deployed as geographically close as possible to the file server to avoid a lot of firewalls in between the scanner and the fileserver. For instance, the scanner can be in the same subnet as the fileserver. In case there are multiple fileservers on a site in different subnets, there is enough license to deploy one scanner for each fileserver. However, it is best to first install one scanner and test if it can scan multiple fileservers on a site. If the scanner does not work properly, it is suggested to install 1 scanner for each fileserver. There is enough license to install as many scanners as needed.

Therefore, the author decided to install the scanner in a selected-by-IT subnet and planned to use the same scanner for all the file servers on the same site. Instead, if there is a problem in the scan speed, the author will request for installation of other windows server hosts for other scanners closer to each fileserver. For this study, only one file server is covered.

6.3 Preparation

6.3.1 Endpoints

Lessons learnt: One of the limitations in large organizations is the latency in the approvals of changes by manager. The change in the protection labels are needed to be approved by the head of security in the subject organization. Planning the change in the labels initiated at this phase. However, this approval was long lasting. It is recommended by the author to consider this fact in advance.

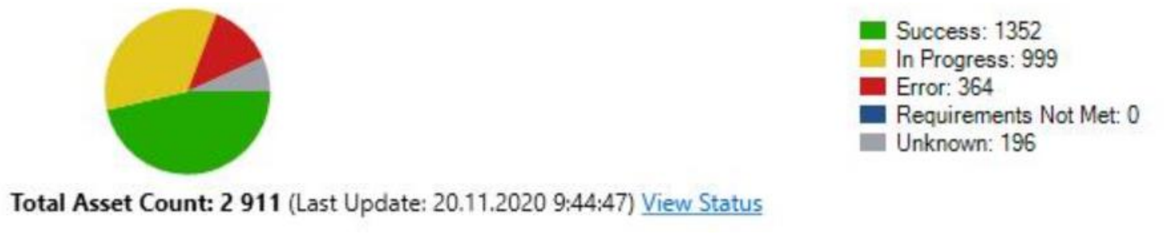
At this point, IT department was contacted to start installing the Unified Labeling agent on the endpoints. It can be either by creating a ticket via the internal ticketing system, or via email or any other formal medium.

Lessons learnt: In case of organizations that manage all the windows machines by System Center Configuration Management (SCCM), this action takes very small amount of time. In case of a large organization that manages some sites with SCCM and some others with other tools, which is the case of this study, it takes a long time to install the agent on about 5000 endpoint machines. Local IT department must contact other geographical IT persons in charge to require such installation which is time consuming.

Lessons learnt: The security specialist to have visibility on the status of the installation progress must be either informed by IT who is responsible for installations via some means such as email, or informed by accessing the tool that IT uses for checking the installation progress. For the case of this thesis, the progress is shown in Figures 22-24 from first time check with IT in SCCM and Service Now Inventory tools.

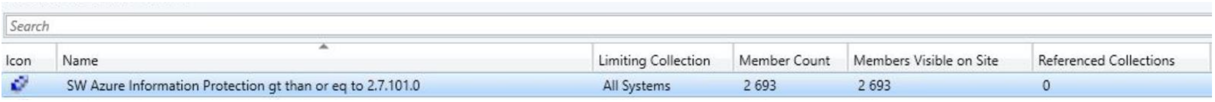
As illustrated in Figure 22 total asset is 2911 managed by SCCM (showed in SCCM tool) while in Figure 24 the installations are 2934 which is more than the total number of the assets. The author decided to rely on SCCM Inventory since it directly receives data from SCCM which is installing the agent.

Since there are roughly 5000 Windows 10 machines involved and in the scope of the installation, the other approximately 2000 machines should be contacted with local IT departments of each site and asked for installations. This step takes a very long time since different focal points are involved.



Total Asset Count: 2 911 (Last Update: 20.11.2020 9:44:47) View Status

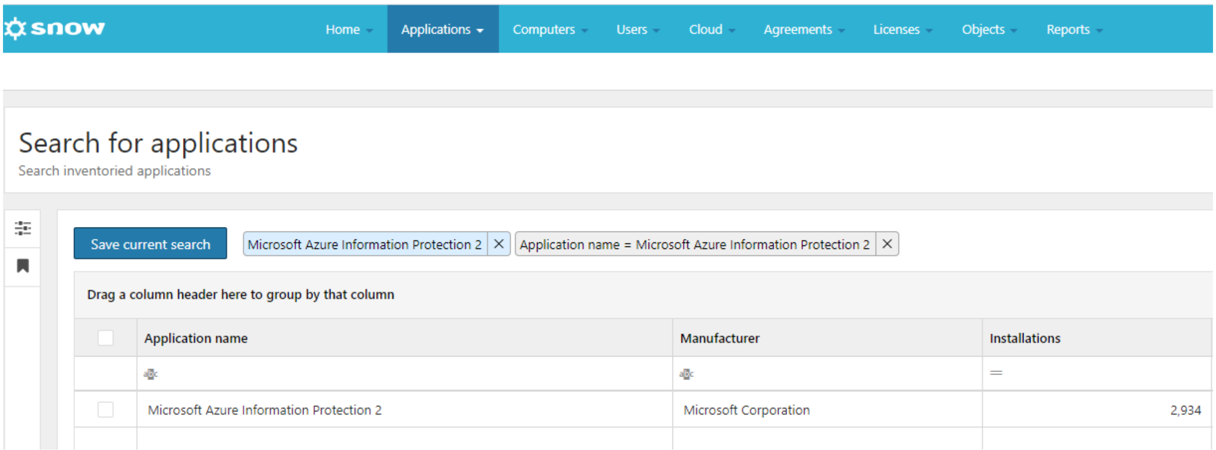
Figure 22. First report of agent installation SCCM.



Icon	Name	Limiting Collection	Member Count	Members Visible on Site	Referenced Collections
	SW Azure Information Protection gt than or eq to 2.7.101.0	All Systems	2 693	2 693	0

Figure 23. Second report of agent installation SCCM.

Lessons learnt: Generally, every action that must be taken by other participants, should be called way ahead of the time of the action. Even informing all the parties by kick-off meeting, it is a fact that everyone in a large organization have their own planned tasks and a sudden task delivered to them, if not urgent, will be their last priority.



Search for applications

Search inventoried applications

Microsoft Azure Information Protection 2 ✕ Application name = Microsoft Azure Information Protection 2 ✕

Drag a column header here to group by that column

□	Application name	Manufacturer	Installations
□	Microsoft Azure Information Protection 2	Microsoft Corporation	2,934

Figure 24. The most recent report of the installation progress (Service Now inventory)

As mentioned earlier, this installation will not stop other phases since the first version had already been installed across the whole company and the logs are continuously being sent to AIP anyway. This installation is an upgrade to the former version. Otherwise, the other steps are independent from this step.

Lessons learnt: Product specific preparations

As explained earlier, Microsoft had announced the deprecation of classic client and had called for the installation of Unified labeling agent on the endpoints as well as activation of Unified Labeling feature that appears as a single button in AIP portal.

Even though this thesis concentrates on the general process of DLP implementation, the author has decided to point out this specific product limitation that was unpredicted and had a negative effect on the timelines of the implementation.

Unified Labeling activation unifies and synchronizes the labels in the two portals of SCC and AIP as explained in Chapter 2 in more details.

The action of activating Unified labeling which apparently is a single button is possible to take a very long time to be done. Specially in case there are already labels in AIP. The best situation is when there are 0 labels in AIP and Unified Labeling becomes activated really fast and without obstacles. However, if there are already labels that are created some time ago and they are using protection setting, it becomes problematic which needs vendor’s support to resolve it.

6.3.2 On-premise

The installation was challenging in terms of permissions. Firstly, some organizations limit internet connection to VMs according to their internal policies. Secondly, the action of installation which was running the script was challenging because the scanner resisted being installed with error below which was related to not having enough permission to install the scanner.

```
[2000008;reason=""The token contains no permissions, or permissions can not be understood."";error_category=""invalid_grant""]" mip::PolicyEngineManagerImpl
permissions are greyed out and company can not be selected to grant permissions
```

The problem was that the installation required the user to have administrative permissions which was also not allowed, and the author had to schedule a call with a system administrator to grant the permission, install the scanner, and then cancel the permission.

The author investigated to troubleshoot the other issues by checking the logs created by the installation script.

6.4 Discovery and classification

6.4.1 Endpoints

Manual labeling of documents (not sensitive, but important)[1]

Manual labeling can be done by the end-users on any document (either sensitive containing document, or confidential organizational document, etc.). It was already mentioned that the validation of the endpoint DLP improvement can be by analyzing the changes in the user

[1] This includes any document that can be labeled as public, confidential, highly confidential because they contain important data such as business data but not sensitive data.

activities from log analytics and Activity log before and after the training on the protection labels.

In this section, the analysis of the data collected before and after the annual information security training on the protection labels to the employees is explained. The employees should be trained how to choose and use the labels. The annual information security training was opened obligatory for 20 days. 3 days after the training the Activity log was analyzed determine if the users could use the labels on their documents with the knowledge that they had gained in the training.

The data was collected from Activity Log and Log Analytics tools provided by Microsoft. The filtering option of date was used. The last 10000 logs of three selected days before and the last 10000 logs after the training were scoped for analysis if the training has had a positive effect on the user behavior in terms of labeling documents. Before the training, the number of the files labeled (all the varied labels) by the users were 2734 out of 10000. After the training, this number increased to 3956. The progress in the user behavior towards labeling their documents is 3956-2733=1223 which is a good progress. The author also took other 10000 scoped files and filtered out the labeled documents to assure of the progress. Without any exception, they all indicated more continuous caring from the users' side in labeling their documents.

The author suggests annual Information Security Training since there is a possibility that either the employee minds less to protect documents over time, or there are newcomers to the organization that are not aware of the labels.

6.4.2 On-premise

On-premise discovery was conducted to discover, Classify, and protect sensitive data. Using any product for improving DLP, it is more appropriate to start with a small scope of one small file server. The first scope for this paper is 3 file repositories in one file server on 1 geographical site. The analysis to identify the properness of the product in use will be explained in this section. Second scan was conducted on 31 repositories on the subject file server including the 3 repositories in the first scan.

For each environment, there must be some analysis conducted by the practitioner to prove that the Product is trustworthy with low number of false positives, or if high, there is a method to reduce the false positives. False positive (FP) is referred to wrong detection of sensitive data.

For example, a document is detected by the scanner as containing sensitive data, while the document does not contain sensitive data when opened and manually checked. True Positive (TP) is the detections that are correct. For example, a file contains sensitive data and it becomes detected correctly. False Negatives (FN) on the other hand, are the files that actually contain sensitive data, while the scanner misses detecting them. The number of the FN is also important as it shows the accuracy and strength of the scanner in detecting sensitive data containing documents correctly and accurately. True Negative (TN) refers to when the scanner correctly outputs that a document does not contain sensitive data.

AIP scanner has this capability that assigns two variables called count number, and confidence to each activity. With these variables, the scanner acknowledges the importance of the file. Count number refers to the number of the sensitive data sets that are detected in a file and confidence which is shown as percentage, refers to how certain the scanner is about the correctness of the detection. For example, count number of 10 and confidence number of 85% means that the scanner speaks */I am 85% sure that this document contains 10 number of sensitive data/.*

It is also important that when large amount of sensitive data is being retrieved, copied, printed or done any activity on, some alerts become sent to a specialist so that the file becomes analyzed. That is the reason why the number of the True positives is also important. The alerts must be in a fair number of true detections daily. Otherwise, the specialists that receive the alerts have to deal with a lot of alerts which are mostly not highly important.

The author proposed the success criteria, accepted by the direct manager, for the discovery and classification phases for on-premise environment is as below:

FP and TP

- Less than 30% of FP. FP/Total detected sensitive data containing documents, and consequently, more than 70% of TP.

In case the above is not met, the existence of a solution to reduce false detections is needed. The author used the two variables of confidence equal to and more than 85% and count number more than 1 to reduce the false positives.

- Less number of alerts[1] only for the documents that contain more than 1 sensitive data and their confidence is more than 80%.

FN and TN

- FN less than 20% out of the total amount of the files and consequently, more than 80% of TN.

First successful scan

The scan was conducted on 3 repositories which discovered 40335 files. Out of those scanned files, there were 27 files that were detected as containing sensitive data (Figure 25). The scan itself discovers all the files regardless of count number or confidence. After the scan is finished, then there are filtering options of count number and confidence.

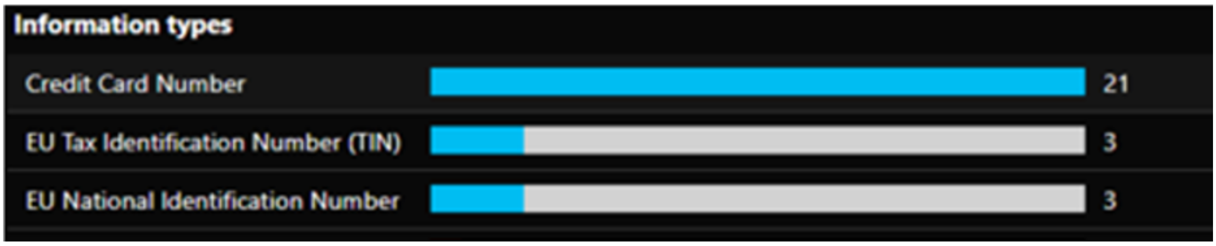

Information types
Credit Card Number
21
EU Tax Identification Number (TIN)
3
EU National Identification Number
3

Figure 25. First scan sensitive data types and the number of the files containing them detected.

The count of more than 1 sensitive data was because some documents are purposed for trainings or instructions and use only 1 sensitive data for the purpose of giving examples. Also, some employees store their own personal data on their computer. excluding these types of sensitive data containing documents, is the best way to prioritize protection in a large organization with huge number of documents that are difficult to handle. Another reason is that companies, if have data breach, will not be fined if small number of sensitive data is breached [47].

The number of FP and TP are comparatively simpler to find. 27 documents that were detected to contain sensitive data were analyzed manually to determine if they were detected correctly or not. out of 27 documents, 19 were detected wrongly and 8 were detected correctly.

[1] The reason to reduce the alerts is to prioritize more important files for the security administrators to focus more on the files that are more critical than the others first. This prioritization contributes to faster action in times when an unaware employee exports a lot of sensitive data from a tool that stores them and saves the file on their computer without protecting them. On the other hand, over protection can be prevented by only protecting the files that have more number of sensitive data.

To identify the number of FN and TN it is required to manually check all the files that are detected not having sensitive data. Even with only 3 repositories, 40335 files are very large scope to manually inspected for FN. Not to mention that some files are very large ones like this very paper. Therefore, to address the number of false negatives, the author started crawling through folders and files manually using a ground truth data set.

Each repository's folders were selected as a categorical folder[1] for analysis. The names of the files as well as the content of 3-4 files were opened and the content was checked manually. Most of the files in one specific folder were having the same format. For example, the pictures of the employees for creating organizational identity card were all stored in one folder. It was mostly enough to even open 1-2 files to understand that a folder is possible to have sensitive data in the opened folder or not. The author did not find any false negative in 2 out of 3 repositories. The reason why the sensitive data containing documents were short in number was that all three repositories where mostly containing images, memorials, office photos, etc, and the scanner does not scan images' extensions for sensitive data. The scanner simply counts them in total but does not inspect them. In total, the author inspected 60 files manually and randomly from the 3 repositories (for each repository 20 files were checked thoroughly) to identify false negatives in a smaller scope and expand the result to the total number of the files in all three repositories. The files mostly found either not supported extension of images by the scanner, or not really containing sensitive data. One folder though was found full of Curriculum Vitae (CV) belonging the applicants for organization's open positions stored by HR. Resumes are considered to be sensitive and they need both protection and retention [48]. The author considered them as false negatives. The folder containing CVs contained 856 files and the whole repository contained 15041 files. Comparatively, 1.13 out of 20 was found as false negatives. Since there were no FN in the other two repositories, 1.3 FN out of 60 were identified.

For better visibility of the scanner performance findings of the first scan, confusion matrix is provided. Confusion matrix is a table containing 4 different combination of predicted and actual value. Predicted value is the result that is expected from a machine (in this context, scanner). For example, a file that contains sensitive data is expected to be detected as containing sensitive data and a file that does not contain sensitive data is expected to be detected as not containing

[1] For example, a folder that is named HR is most probably related to Human Resource department. So, this folder is the category of HR.

sensitive data. Predicted values are positive and negative. Actual value is the result that the machine (scanner) shows. The scanner either shows that a file contains sensitive data (True), or a file does not contain sensitive data (False). The combination of predicted and actual data gives the information about if the scanner is correctly detecting or not. Confusion matrix Table 3 indicates the findings related to the first scan.

Table 3. Confusion Matrix first scan.

Confusion Matrix Total files 40335		**Predicted Value**	
		P	N
Actual Value	T	TP= 8	TN= 39,435
	F	FP= 19	FN= 873

- Predicted Value: The value that is expected from the scanner to detect.

- Actual Value: The Value that the scanner detects in practice.

The accuracy of the scan is calculated by below metrics:

Accuracy[1]: (TP+TN)/(TP+FP+FN+TN) =39,443/40335 ~ 0.97

Recall[2]: TP/(TP+FN) = 8/ (8+873) ~ 0.009

Precision[3]: TP/(TP+FP) = 8/ (8+19) ~ 0.29

FP and TP

- As shown in the confusion matrix, the number of FP out of all the detected sensitive data containing documents is 19 out of 27 which is more than 70%. Thus, this result refuses the success criteria of FP less than 30%. Then, it is required to reduce the number by conditions below:

Confidence more than 85% and count more than 1 to reduce false positives. False positives were reduced from 19 out of 27 detected sensitive data containing documents to 1 out of 27

[1] The number of the correct detections out of total number of the files.

[2] The number of sensitive data containing documents detected out of the number of the real sensitive data containing documents existed on the file server.

[3] The number of the correct detected files with sensitive data out of all the detections of sensitive data.

detected sensitive data containing documents which reduces FP to less than 30% defined in success criteria.

- Confidence more than 80% and count more than 1 to reduce alerts.

Alerts reduced to 4 out of 27 detected sensitive data containing documents. The reason for reducing the true positives are to use the conditions related to them for creating alerts. From administrative point of view, less alerts and less false positives are two criteria that are important. Alerts can lead the administrators to apply protection labels on the documents that contain sensitive data using the scanner in the protection mode.

FN and TN

The number of the FN which is ~2.16%, met the success criteria of less than 20%. Consequently, the number of TN which is ~ 97.84%, is more than 80% and meets the success criteria.

The Decisions based on the result varies from organization to organization based on the expectations. However, the decisions in the subject organization was to approve expanding the process and installations of the scanner on other geographical sites for data protection purposes which are out of the scope of this study.

The success criteria about the FN, FP, TP and TN were met in a small scope of 3 file repositories. The increase of count number more than 1 and confidence of more than 85 and 80 (FP and alerts respectively) also provided a solution to reduce both FP and Alerts. The discovery on the 3 repositories in this section then was the basis to decide to proceed with the next step of discovery of the whole file server.

Second successful scan

The initial success criteria apply for the second scan too. The purposes for the second scan are as below:

1. It is important to understand if the scanner works properly in a larger scope without corruptions or errors.
2. Classifying documents as 1) Containing sensitive data and 2) Not containing sensitive data. Also, classifying the documents with sensitive data by type of sensitive data such as credit card number, EU phone number, etc similar to the figure 25.

3. Inspection and discovery of the whole file server as the basis to take action on applying protection.

This scan was conducted on 31 repositories which scanned 540335 documents. Figure 26 illustrates only highest number of sensitive information types that were detected. The list in the figure continues with other sensitive information types which are cropped in the figure 26. There were in total 1943 files out of 540335 files which were detected as containing sensitive data.

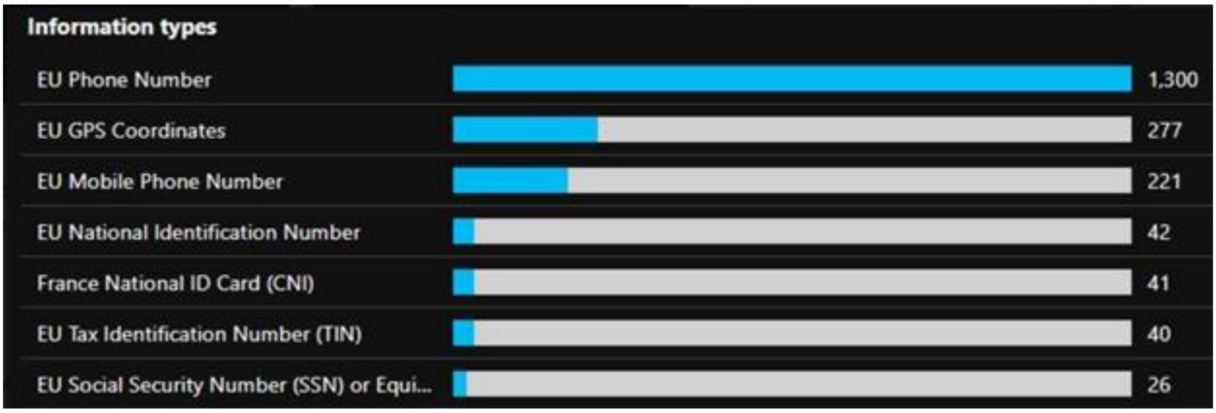


Figure 26. Second scan sensitive data types and the number of the files containing them detected.

Since the number of the detected files that contained EU phone number was very large and the author needed to open each file log and analyze the false positives, a scope of 50 files as a ground truth dataset was analyzed from all the 1300 files. (See Figure 26) EU Global Positioning System (GPS) coordinates and EU mobile phone number were skipped (the choice can be based on organizational requirements). Others were all analyzed to calculate the number of the False positives and true positives. In total, 240 files were analyzed, and all the result was expanded to the whole scope.

The numbers of the FP, FN, TP, and TN are illustrated in Table 4.

FP and TP

- As shown in the confusion matrix, the number of FP out of all the detected sensitive data containing documents is 1263 which is 65%. Thus, this result refuses the success criteria of FP less than 30%. Then, it is required to reduce the number by conditions below:

Confidence more than 85% and count more than 1 to reduced false positives. False positives were reduced from 1263 detected sensitive data containing documents to 364 detected sensitive data containing documents which reduces FP to 18% which is less than 30% defined in success criteria.

- Confidence more than 80% and count more than 1 was used to reduce alerts. Alerts reduced to 170 out of 1943 detected sensitive data containing documents. 170 TP means that when filtered by confidence of more than 80% and count of more than 1, Alerts will be reduced from 680 to 170 which is from 65% to 8%.

Table 4. Confusion matrix second scan.

Confusion Matrix **Total files 540335**		**Predicted Value**	
		P	N
Actual Value	T	TP= 680	TN=512,341
	F	FP= 1,263	FN=26,051

The accuracy of the scan is calculated by below metrics:

Accuracy: (TP+TN)/(TP+FP+FN+TN) = 513,021/ 540335 ~ 0.94

Recall: TP/(TP+FN) = 680 / (680+26,051) ~ 0.025

Precision: TP/(TP+FP) = 680/ (680+1263) ~ 0.34

FN and TN

To estimate the number of the false negatives, the author proceeded with the same method of the first successful scan data. However, this time the ground truth data set is in a larger scope of a greater number of repositories (31 including the repositories from the first successful scan), folders, and files (540335). Based on the names of the folders, the author divided the file server into categories of files.

In each repository, 10 documents were inspected manually which in total is 310 documents. Also, 3 of those 31 repositories were inspected earlier in the first scan and were not required to be inspected a second time. The author found some files in different categories to contain sensitive data.

- IT related files containing IP addresses as well as the assets names related to them.
- HR related Forms filled with personal data
- Another folder named recruiter containing resumes of applicants for open positions.
- A folder containing the annual salary of all the site's employees in different years, address, marital status which identified a single person (PII).
- Employee's passport or visa scans (even though the author knows that the tool does not inspect images, these files were very important, and they were considered as false negative to be reported and protected accordingly by some other means)

The number of the false negatives in the scope of 310 files was 15. The number of the FN which is ~ 4.8%, met the success criteria of less than 20%. Consequently, the number of TN which is ~ 95.2%, is more than 80% and meets the success criteria.

6.5 Protection

6.5.1 Endpoints

Automatic/Recommending labeling for sensitive data

Automatic label was not approved to be applied and published for the end-users by management level. The recommendation was orally promised from the author's direct manager to be approved for deployment though. Nevertheless, it is still possible to analyze and predict the results of using Automatic/ Recommendation labeling in terms of DLP improvement. Automatic/Recommending a label was initially proposed by the author to solve the problem of forgetting to label by the end-users that was initially identified in the questionnaire. Automatic/recommending labeling can be used only for the documents containing sensitive data which can be identified by Pre-defined templates.

Automatic labeling is to force push label a document without user interaction and in case the user decides to change the label to a lower protecting label, the user can be forced to provide justification or not. Recommending a label feature detects the sensitive data in documents but does not force apply a label on it. It only provides a pop-up to both inform and notify the user of the existing sensitive data and reminding the user to apply a label on the sensitive containing document.

Nevertheless, the author attempted to analyze the progress that can be made in DLP by automatically applying labels on sensitive data containing documents.

10000 documents were scoped in the logs using Activity log and Log analytics provided by Microsoft. These logs, as explained earlier are collected from the AIP clients (no matter classic or UL, all the logs become collected in Activity log[1]) installed on the endpoints. Out of those, 5436 documents were detected as containing sensitive data. Another finding was that only 42 files out of 5436 files were labeled and only 2 out of these files were protected (by user restriction and encryption) by the end-users.

The author used the finding in conditioning for on-premise to reduce the false positives. If 85% confidence and count more than 1 is chosen for filtering, about 475 files should be considered as false positives based on below calculations:

21/240 = x/5436 => x= 475

4961 files are not false positives. Out of those of NOT false positives, to reduce the alerts, the true positives were reduced to 85% confident and more than 1 count number. 1019 files with confidence of 85% and count of more than 1 were identified as true positive and must be in high priority to be protected but they are not. The calculations below are using the results in endpoints analysis for alert reduction.

45/240= x/5436 => x= 1019

comparing to all the scoped files, 10000, the number 1019 sensitive data with confidence more than 85% and count more than 1 is about 10% of the scoped files.

In conclusion, up to 10% of documents in the subject organization are identified as containing sensitive data that had remained unprotected and must have been labeled and protected by encryption. If automatic labeling is applied, these 10% files containing sensitive data will become automatically labeled and protected and consequently, DLP will be improved.

The analysis showed that 10% of the total 10000 files are sensitive data with confidence of more than 85 percent and count of more than 1 which would be automatically protected when

[1] This means even if half of the endpoints have classic client and the other half have UL client, all the logs will be gathered and illustrated in the activity log.

applied. This is a major improvement in DLP. In the case of this paper, the management decision was to deploy recommending labels to the end-users rather than force pushing the labels for their documents. This is the same mechanism of detection for automatic labeling but the slight difference is that instead of labeling a document automatically, the user will be prompted with a recommended label and she/he has the choice to whether apply the label, or skip it with a justification. Justifications will also be visible in the Activity log.

6.5.2 On-premise

In case both modes of discovery and protection is used on a file server, there will be experience gained from the protection mode as well as discovery mode. For the case of this study, the management preferred to only apply discovery mode on the file server.

On-premise protection from the back end administrative side will only be applied on sensitive data containing documents and not the documents that are important or confidential but do not contain sensitive data. Therefore, protection in on-premise environment will only be applied on sensitive data containing documents.

As in the analysis of discovery and classification was explained, 680 out of 1943 files were true positives. However, it is only the organizational decision to support applying protection on all these 1943 files, or only apply protection on the detections leading alerts. Leading alerts means that the organization can choose to apply protection on all the sensitive data containing documents or reduce protection on only the documents that contain more than 1 sensitive data or more confidence. In either preference, DLP will be improved by protecting sensitive containing files. In the first step that a small scope of 3 repositories were scanned, the conditions that were applied reduced the false positives from 19 out of 27 (70%) to 1 out of 27 (3%) and reduced the number of Alerts from 8 out of 27 (29%) to 4 out of 27 (14%).

The result provided more real detection and less alerts that notifies the security analyst only when there is a confidence of more than 80% that more than 1 sensitive data strings existed in a document. This helps optimizing the detections and alerts which contribute to faster performance of security analysts in analyzing and protecting the sensitive data containing documents. Reducing the alerts can be a solution to prioritize protection, react faster to protect more important files, and reduce over protection. This is a major progress in the improvement of DLP for on-premise environment. The discovery result can also be utilized in determining retention labels for very old files that can only be detected by the on-premise AIP scanner and not the endpoint client.

The management decision was totally positive about the findings of discovery and approved proceeding to the next level of installing the scanner on other major sites which is out of the scope of this paper.

7 Chapter 7 – Conclusion

This study proposes and follows a detailed process of improving Data Loss Prevention in a large organization. The objective is to prove if the proposed detailed process improves the status of DLP in the subject organization. The scope selected was two environments of on-premise and endpoints. The Process proposed was deployed on a real organization using action research methodology as the main methodology for the whole process. DLP status of both environments of on-premise proved to be improved by the process deployed and analyzed for accuracy and benefits.

The process starts with Initiation, and continues with Planning, Preparation, discovery, classification and Protection phases. The last phase, which is monitoring, and it is a continuous phase, was not covered in the current thesis.

Each phase of the process repeated for each environment of on-premise and endpoints was explained and analyzed with a specific methodology.

For Endpoints, the coverage is protecting both the files that contain sensitive data (automatic labeling), and the important files that need protection but do not contain sensitive data (manual labeling). These important files contain internal business data which must also be protected. The process starts with identifying the context problems as well as conducting a questionnaire to identify where sensitive data resides and what prevents the user to apply labels as well as deciding about the optimization of the labels based on end user preferences. To prepare for the implementation, the agent that collects and sends user activity logs to the Azure portal must be installed. This installation duration was long in practice as some sites where supported by SCCM and some were not. However, the next phases of discovery, classification and protection continued since they could be done in parallel with the agent installation (as mentioned earlier, the classic client had already been installed on all the endpoints and the logs are already collected from the endpoints for discovery and classification. This new installation of the UL agent is only an upgrade to classic client). The logs appear on Azure portal which were analyzed before and after annual information security training. The analysis showed that user behavior improves the DLP on the endpoint significantly from 2734 files protected to 3956 files protected out of 10000 scoped files.

For on-premise, the process starts with recognition of the context as well as the problems to solve within the context followed by the consultancies with relevant focal points to planning the implementation. The preparations for installing the scanner including the host server installation as well as the product related application registration and account token was done at this point. This was followed by installing the scanner and conducting a scan to discover and classify the data (sensitive and not sensitive data containing documents) as well as analyzing the logs to identify the FP,FN,TP,TN as well as finding a method to reduce false positives as well as reducing alerts related to true positives. TP detections were then decided to be protected. The number of the FN which was ~ 4.8%, met the success criteria of less than 20%. Consequently, the number of TN which is ~ 95.2%, is more than 80% and meets the success criteria. The number of the FP was rather high 65% which was reduced by applying conditions of confidence 85 and count more than 1 to 18%. Reducing alerts was applied by TP confidence number of more than 80 and the count number more than 1 was preferred to be set. The conditions of confidence and count can vary based on organizational preference in how many sensitive data strings should be detected in a file to trigger alerts. The result showed that the FP number reduced with the conditions applied and the alerts reduced to more critical alerts instead of alert for every single TP detection. Protecting the TP detected files improves the DLP status of the organization.

7.1 Lessons learnt

Lessons learnt 1: One of the most valuable pieces of information in the result was that the author was mostly searching for the departments that deal with sensitive data initially. However, after the questionnaire responds were collected, the author found that in one department, some employees deal with sensitive data and some do not. This means that dealing with sensitive data is not only related to the department in which the employee works, but also the position title of the employee can represent their connection with sensitive data.

Lessons learnt 2: The same result for uploading a labeled document on Sharepoint was applied and resulted for Onedrive. The difference between Onedrive/SharePoint and office applications of Excel, Word, PowerPoint and outlook is that the labels do not appear in Onedrive/SharePoint. However, if a file is created on desktop and uploaded on Onedrive/SharePoint, the label travels with the document and protects it. The problem was that if a document is created directly online from Onedrive/SharePoint, unfortunately they cannot be protected from there since the labels do not appear in Onedrive/SharePoint.

Lessons learnt 3: In a large organization, there are varied number of file servers on each site. Some sites have only one fileserver and some more than one. In case there is only one fileserver on a site, one scanner can be deployed as geographically close as possible to the file server to avoid a lot of firewalls in between the scanner and the fileserver. For instance, the scanner can be in the same subnet as the fileserver. In case there are multiple fileservers on a site in different subnets, there is enough license to deploy one scanner for each fileserver. However, it is best to first install one scanner and test if it can scan multiple fileservers on a site. If the scanner does not work properly, it is suggested to install 1 scanner for each fileserver. There is enough license to install as many scanners as needed.

Therefore, the author decided to install the scanner in a selected-by-IT subnet and planned to use the same scanner for all the file servers on the same site. Instead, if there is a problem in the scan speed, the author will request for installation of other windows server hosts for other scanners closer to each fileserver.

Lessons learnt 4: One of the limitations in large organizations is the latency in the approvals of changes by managers. The change/creation of the protection labels are needed to be approved by the head of security in the subject organization. It takes almost half a year to gain the approvals. It is recommended by the author to consider this fact way in advance.

Lessons learnt 5: In case of organizations that manage all the windows machines by System Center Configuration Management (SCCM), the action of installing the UL client takes very small amount of time. In case of a large organization that manages some sites with SCCM and some others with other tools, which is the case of this study, it takes a long time to install the agent on about 5000 endpoint machines. Local IT department must contact other geographical IT persons in charge to require such installation which is time consuming.

Lessons learnt 6: The security specialist to have visibility on the status of the agent installation progress must be either informed by IT who is responsible for installations via some means such as email, or informed by accessing the tool that IT uses for checking the installation progress.

Lessons learnt 7: For on-premise, in case in the first scan attempt, both modes of discovery and protection is used, there will be more experience gained from the protection mode as well as classification mode.

Lessons learnt 8: One of the major limitations of the study was the fact that identifying the number of the false negatives can be very difficult to be done manually when large number of files are involved. The utilization of ground truth dataset as a representative of all the files to analyze the number of the false negatives is suggested by the author as a method to estimate the number of the false negatives.

7.2 Further study

In a big picture of the organization, the whole DLP alerts will be handed over to Security Operation Center (SOC). Further study can be conducted on Incident Response Plan (IRP) following the investigation of the alerts that are more critical related to the files that contain large amount of sensitive data and if exposed, would create damage to the company.

The author also suggests a further article research around monitoring and taking corrective actions on the possible problems such as confusions from end users' side (informing administrator by report an issue button), misconfigurations of labels, and availability problems when restrictions are applied that follows the whole process of this study.

Another scope for further study is DLP improvement in cloud environment which was not covered in this study. The fact that the files should carry their protection everywhere they reside, can lead to integrations with other tools related to cloud environment, or utilizing the tools provided by the same vendor for this study's product which is Microsoft.

In this study, departmentalization of the labels was not applied since dealing with sensitive data deferred within same department for different positions. However, in case a practitioner is applying labels on a small organization, label taxonomy based on the departments or in case the sensitive information is concentrated in one or two departments, it would be much simpler to configure labels specific to each department which only appears for them. Further study can be conducted for designing labels for a small organization and departmentalization of the labels.

A very interesting further study can relate to psychological aspect of DLP. This includes the fact that some employees do not mind protecting their documents. The reasons why the employees do not mind protecting documents as well as how to reason the importance of DLP for the employees to draw their attention and increase their motivation to contribute for better results in DLP improvement can be covered in further research.

Automatic detection of the confidential information in practice was not covered in this study. This can be covered in the future study using inspection methods of text mining to identify the keyword appearance frequency [39] in the documents and analyzing the results in a real environment practice.

References

[1] I. Beerepoot, I. van de Weerd, and H. A. Reijers, *Business Process Improvement Activities: Differences in Organizational Size, Culture, and Resources*, vol. 11675 LNCS. Springer International Publishing, 2019.

[2] G. Garrison, “An assessment of organizational size and sense and response capability on the early adoption of disruptive technology,” *Computers in Human Behavior*, vol. 25, no. 2. pp. 444–449, 2009, doi: 10.1016/j.chb.2008.10.007.

[3] “@ www.keepnetlabs.com.” [Online]. Available: https://www.keepnetlabs.com/the-biggest-data-breaches-in-the-first-half-of-2020/.

[4] “information-protection @ docs.microsoft.com.” [Online]. Available: https://docs.microsoft.com/en-us/microsoft-365/compliance/information-protection?view=o365-worldwide.

[5] M. Hart, P. Manadhata, and R. Johnson, “Text classification for data loss prevention,” *HP Lab. Tech. Rep.*, no. 114, pp. 1–21, 2011.

[6] P. Raman, H. G. H. Kayacık, and A. Somayaji, “Understanding Data Leak Prevention,” *Annu. Symp. Inf. Assur.*, vol. 2016, no. 3, pp. 27–31, 2011, doi: 10.1109/IConAC.2015.7313979.

[7] H. Alkilani, M. Nasereddin, A. Hadi, and S. Tedmori, “Data exfiltration techniques and data loss prevention system,” *Proc. - 2019 Int. Arab Conf. Inf. Technol. ACIT 2019*, pp. 124–127, 2019, doi: 10.1109/ACIT47987.2019.8991131.

[8] “what-personal-data-considered-sensitive_en @ ec.europa.eu.” [Online]. Available: https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_en.

[9] “@ www.statista.com.” [Online]. Available: https://www.statista.com/statistics/996456/data-breaches-reported-in-europe-by-country/.

[10] Consult Hyperion, “GDPR: banks, breaches and billion euro fines,” no. June, 2017.

[11] “@ www.statista.com.” [Online]. Available: https://www.statista.com/statistics/273575/average-organizational-cost-incurred-by-a-data-breach/.

[12] “@ www.comparitech.com.” [Online]. Available: https://www.comparitech.com/net-admin/data-loss-prevention-tools-software/.

[13] “quickstart-label-dnf-protectedemail @ docs.microsoft.com.” [Online]. Available: https://docs.microsoft.com/en-us/azure/information-protection/quickstart-label-dnf-protectedemail.

[14] “configure-policy @ docs.microsoft.com.” [Online]. Available: https://docs.microsoft.com/en-us/azure/information-protection/configure-policy.

[15] “what-is-information-protection @ docs.microsoft.com.” [Online]. Available: https://docs.microsoft.com/en-us/azure/information-protection/what-is-information-protection.

[16] “office-365-securitycompliance-center @ docs.microsoft.com.” [Online]. Available: https://docs.microsoft.com/en-us/office365/servicedescriptions/office-365-platform-service-description/office-365-securitycompliance-center.

[17] “client-version-release-history @ docs.microsoft.com.” [Online]. Available: https://docs.microsoft.com/en-us/azure/information-protection/rms-client/client-version-release-history.

[18] O. Akinrolabu, I. Agrafiotis, and A. Erola, “The challenge of detecting sophisticated attacks: Insights from SOC analysts,” *ACM Int. Conf. Proceeding Ser.*, 2018, doi: 10.1145/3230833.3233280.

[19] D. Fujs, A. Mihelič, and S. L. R. Vrhovec, “The power of interpretation: Qualitative methods in cybersecurity research,” *ACM Int. Conf. Proceeding Ser.*, 2019, doi: 10.1145/3339252.3341479.

[20] C. T. Berry and R. L. Berry, “An initial assessment of small business risk management approaches for cyber security threats,” *Int. J. Bus. Contin. Risk Manag.*, vol. 8, no. 1, p.

1, 2018, doi: 10.1504/ijbcrm.2018.10011667.

[21] A. Chattopadhyay, D. Christian, A. Ulman, and C. Sawyer, "A Middle-School Case Study: Piloting A Novel Visual Privacy Themed Module for Teaching Societal and Human Security Topics Using Social Media Apps," *Proc. - Front. Educ. Conf. FIE*, vol. 2018-Octob, 2019, doi: 10.1109/FIE.2018.8659278.

[22] R. M. Davison, M. G. Martinsons, and N. Kock, "Principles of canonical action research," *Inf. Syst. J.*, vol. 14, no. 1, pp. 65–86, 2004, doi: 10.1111/j.1365-2575.2004.00162.x.

[23] E. Tüzün, B. Tekinerdogan, Y. Macit, and K. İnce, "Adopting integrated application lifecycle management within a large-scale software company: An action research approach," *J. Syst. Softw.*, vol. 149, pp. 63–82, 2019, doi: 10.1016/j.jss.2018.11.021.

[24] R. L. Baskerville, "Investigating Information Systems with Action Research," *Commun. Assoc. Inf. Syst.*, vol. 2, no. October, 1999, doi: 10.17705/1cais.00219.

[25] U. K. Bamel, S. Rangnekar, R. Rastogi, and S. Kumar, "Organizational process as antecedent of managerial flexibility," *Glob. J. Flex. Syst. Manag.*, vol. 14, no. 1, pp. 3–15, 2013, doi: 10.1007/s40171-013-0026-9.

[26] A. M. Magdaleno, V. T. Nunes, R. M. Araujo, and M. R. S. Borges, "Flexible organizational process deployment," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4402 LNCS, pp. 679–688, 2007, doi: 10.1007/978-3-540-72863-4_69.

[27] "tutorial-dlp @ docs.microsoft.com." [Online]. Available: https://docs.microsoft.com/en-us/cloud-app-security/tutorial-dlp.

[28] D. L. Prevention, "NS Ins titu te Au tho r r eta ins ful l rig," 2020.

[29] S. Zawoad and R. Hasan, "FECloud: A Trustworthy Forensics-Enabled Cloud Architecture," *Adv. Digit. Forensics XI*, no. 462, pp. 271–285, 2015, doi: 10.1007/978-3-319-24123-4.

[30] "index @ www.endpointprotector.com." [Online]. Available: https://www.endpointprotector.com/.

[31] “@ www.trustwave.com.” [Online]. Available: https://www.trustwave.com/en-us/.

[32] “index @ mydlp.com.” [Online]. Available: https://mydlp.com/.

[33] “@ www.helpnetsecurity.com.” [Online]. Available: https://www.helpnetsecurity.com/2010/05/03/opendlp-data-loss-prevention-tool/#:~:text=OpenDLP is a free and,from a centralized web application.

[34] G. Lopez, N. Richardson, and J. Carvajal, “Methodology for data loss prevention technology evaluation for protecting sensitive information,” *Rev. Politécnica*, vol. 36, no. 3, p. 69, 2015.

[35] C. Mercy Praba, G. Satyavathy, C. Mercy, P. M. P. Scholar, and G. Satyavathy, “A Technical Review on Data Leakage Detection and Prevention Approaches,” *J. Netw. Commun. Emerg. Technol.*, vol. 7, no. 9, pp. 67–72, 2017.

[36] K. Kaur, I. Gupta, and A. K. Singh, “A Comparative Evaluation of Data Leakage/Loss Prevention Systems (DLPS),” pp. 87–95, 2017, doi: 10.5121/csit.2017.71008.

[37] N. Rajagopal, K. V. Prasad, M. Shah, and C. Rukstales, “A new data classification methodology to enhance utility data security,” *2014 IEEE PES Innov. Smart Grid Technol. Conf. ISGT 2014*, pp. 14–18, 2014, doi: 10.1109/ISGT.2014.6816451.

[38] H. P. Luhn, “A Statistical Approach to Mechanized Encoding and Searching of Literary Information,” *IBM J. Res. Dev.*, vol. 1, no. 4, pp. 309–317, 2010, doi: 10.1147/rd.14.0309.

[39] E. Alparslan, A. Karahoca, and H. Bahşi, “Security-level classification for confidential documents by using adaptive neuro-fuzzy inference systems,” *Expert Syst.*, vol. 30, no. 3, pp. 233–242, 2013, doi: 10.1111/j.1468-0394.2012.00634.x.

[40] F. F. Moghaddam, M. Yezdanpanah, T. Khodadadi, M. Ahmadi, and M. Eslami, “VDCI: Variable data classification index to ensure data protection in cloud computing environments,” *Proc. - 2014 IEEE Conf. Syst. Process Control. ICSPC 2014*, no. December, pp. 53–57, 2014, doi: 10.1109/SPC.2014.7086229.

[41] R. Singh, V. Mindel, and L. Mathiassen, “IT-based revenue cycle management: An action research into relational coordination,” *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*,

vol. 2016-March, pp. 3152–3161, 2016, doi: 10.1109/HICSS.2016.396.

[42] K. Olesen and M. D. Myers, “Trying to improve communication and collaboration with information technology:An action research project which failed,” *Inf. Technol. People*, vol. 12, no. 4, pp. 317–332, 1999, doi: 10.1108/09593849910301621.

[43] É. St-Pierre, C. Boton, and G. Lefebvre, “Implementing Mobile Technology on Construction Sites: An Ethnographic Action Research Approach,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11792 LNCS, pp. 142–150, 2019, doi: 10.1007/978-3-030-30949-7_16.

[44] “221792 @ techcommunity.microsoft.com.” [Online]. Available: https://techcommunity.microsoft.com/t5/microsoft-security-and/installation-configuration-and-usage-of-the-aip-scanner/ba-p/221792.

[45] “index @ www.meta-chart.com.” [Online]. Available: https://www.meta-chart.com/.

[46] “client-admin-guide @ docs.microsoft.com.” [Online]. Available: https://docs.microsoft.com/en-us/azure/information-protection/rms-client/client-admin-guide#should-you-deploy-the-azure-information-protection-client.

[47] “article-83-general-conditions-for-imposing-administrative-fines-GDPR @ www.privacy-regulation.eu.” [Online]. Available: https://www.privacy-regulation.eu/en/article-83-general-conditions-for-imposing-administrative-fines-GDPR.htm.

[48] “guide-identifying-personally-identifiable-information-pii @ www.technology.pitt.edu.” [Online]. Available: https://www.technology.pitt.edu/help-desk/how-to-documents/guide-identifying-personally-identifiable-information-pii.

[49] “What Is Sensitive Data? Sensitive Data Definition & Types.” [Online]. Available: https://cipherpoint.com/blog/what-is-sensitive-data/

[50] “How to determine the sensitivity of information” [Online]. Available: https://www.spirion.com/blog/how-to-determine-the-sensitivity-of-information/

Appendix 1- Questionnaire content

Target population for a questionnaire survey: All employees of the organization in 2 geographical sites

Number of the participants in the questionnaire: 637 employees

The objectives of the questionnaire:

1. The departments that are directly in contact with sensitive data
2. The number of the employees who are aware of protection labels out of total respondents.
3. The number of the employees who use the protection labels on sensitive containing documents.
4. The availability limitation extension of protection labels.
5. Deciding if the employees need training for using the labels.
6. The number of the employees who are aware of internal related policies.
7. The complains/concerns of the employees about the current status of the labels.
8. The suggestions of employees.
9. Deciding if it is better to have a new set of label taxonomy.
10. Determining if there is a need to have default label, automatic labeling assignment, label reminder prompting.

Type: structured questionnaire filled by the participants

Section titles

The author conducted the 5 items below as section titles of the questionnaire as below:

- 1-3 Job description
- 4-5 Employee DLP knowledge
- 6 Sensitive data dispersion across the company
- 7-8 Employee concerns
- 9-11 Employee requests

//The beginning of the questionnaire//

Job description

1. What is your position title?

2. Which department do you work in?
3. How long have you been working in this company?

- Less than a year
- From 1 to 3 years
- From 3 to 10 years
- More than 10 years

Employee DLP knowledge

4. How well are you familiar with <here the name of the internal policy was mentioned>?

- I know if very well
- Partially (have read the document, but do not remember much)
- I know it exists but have not found time to read it
- It is unknown to me

5. Are you familiar with the Microsoft Word, Excel, PowerPoint, and outlook protection labels?

- Yes
- No

Sensitive data dispersion across the company

In case the answer to the last question is

- *Yes, please continue with the next questions.*
- *No, please skip questions 6 to 9, and continue from question 10.*

6. How often do you deal with sensitive data (sensitive data is defined above)?

- Always (100% of your documents contain sensitive data)
- Often (50%-99% of your documents contain sensitive data)
- Sometimes (20%-49% of your documents contain sensitive data)
- Rarely (0%-19% of your documents contain sensitive data)
- Never (0% of your documents contain sensitive data)

Employee concerns

In case the respondent

- *Did not choose Never for question 6, move to question 7*
- *Chose Never for question 6, continue from question 10*

7. How often do you use Protection labels on documents that contain sensitive data?

- Always
- Often
- Sometimes
- Rarely
- Never

In case the respondent

- *Did not choose always for question 7, continue from question 8.*
- *Chose always for question 7, continue from question 9.*

8. What are the main reasons why you may not protect the documents that contain sensitive data? (you can choose more than one option)

- Weak protection: You believe the protection labels are not strong protection mechanisms.
- No mind: choose this if you do not mind protecting documents.
- Confusion: choose this if the descriptions of the labels are not clear or you believe the number of the labels is high (ex, you think there should be 3 labels instead of 6 labels), or you believe that the names of the labels are not simple.
- Time waste: choose this if you believe understanding and selecting an appropriate label is time consuming or difficult.
- Forgetting: choose this if you usually forget to label your document.
- Availability limitation: choose this if cases happened in your experience that you sent an email or shared a document and the purposed recipient had difficulties doing actions such as opening, editing, printing, copying, etc the file.
- Lack of training: choose this if even though you know the labels, you need to be trained to fully understand when to use what label.
- Other reasons: choose this if you have experienced any other problem on the Employee requests.

9. Do you agree/disagree that each department should have their own specific protection labels that are hidden to other departments? (ex. HR should have a label named employee personal data, etc)?

- Agree
- Disagree

10. Do you need training/additional training to become familiar with protection labels and how to apply them?

- Yes, I need to learn where to find the labels, how to use the labels, and how to choose the best labels for my documents.
- Yes, I need to learn how to use the labels, and how to choose the best labels for my documents
- No, I already know about the labels.

In case the answer to question 10 is

- *Yes, please continue to question 11.*
- *No, do not continue. End and submit the questionnaire directly from this question.*

11. What platform of training do you prefer?

- Video training
- Document training

//The end of the questionnaire//

Appendix 2- Questionnaire banner

Some Additional knowledge before you start:

Definition of sensitive data: It refers to personal data such as name, last name, ID number… as well as medical records and financial records such as credit card number. Sensitivity labels of our company: (for the anonymity and confidentiality reasons, this section has been removed.) Estimated time of reply: 1-5 minutes. The questionnaire will be closed on Mon Oct 5.

Appendix 3- First email content to call the respondents

//Data Loss Prevention Improvement

As a family, hand-in-hand we stand tall against any harm to our company's confidential information and individual's personal data (either the one of employees, customers, etc.). A survey is conducted by Security with the purpose to get an overview of your overall data loss prevention practices in order to take further possible actions towards Data Loss Prevention improvement according to your responses.

Respondents' personal data will neither be collected, nor exposed. There will also not be any judgmental attitude towards your answers to the questions. Your honest replies provide the biggest value to the Security team and it is important that every single employee attends this survey. Please do not google to fill in the survey for us to receive real and reliable information. The questionnaire can be found here (link to the form was placed on the word "here") and it will close on Monday, Oct 5th. The estimated time of reply is 1-5 minutes, so please find this short time to help the Security team.

Best regards and thank you for your help,

Arefeh Kalkhoran

Information Security Engineer

Appendix 4- Second email content to call the respondents

This is to kindly remind you that the Security conducted a survey to get an overview of your overall data loss prevention practices in order to take further possible actions towards Data Loss Prevention improvement according to your responses.

The questionnaire can be found here (The link to the questionnaire was linked to the word "here") and it **will close on Monday, Oct 5th.** The estimated time of reply is only 1-5 minutes, so please find a moment to help the Security team, if you have not done so yet.

Thank you for your help,

Arefeh Kalkhoran

Information Security Engineer

Appendix 5- New Label Taxonomy setting plan

Main Label name	Non-business	Public	General/Other/Internal	Confidential/ Highly-Confidential	Sensitive/Top Secret
Main Label Description	Individual private docs which are not related to the the company	Ads and Global announcements	Default In case the document does not fall into other	Business documents, Internal Policies, Organizational info, Asset info	Sensitive data related to employees, customers or players B2C and licensee B2B. Top secret business related information.

			categories of labels						
Sub-label name	No sub-label	No sub-label	No sub-label	All Employees	Custom Permission	Anyone	Personal data we can have: - personal data Internal--> only internal employees - personal data External --> All employees and external parties	Financial data we can have: - Financial data Internal --> only internal employees - Financial data External --> All employees and external parties	Business Information we can have: - Business Information Internal --> only internal employees - Business Information External --> All employees and external parties
Sub-label Description				If confidential document is sent to internal employees	In case the document creator intends to change the permissions based on her/his need.	If confidential document is sent to external Parties other than internal	Related to employees and customers: “An individual political opinion or party affiliati	Financial information – Credit card numbers, bank account information, and social security numbers. Government information	“Accounting data, trade secrets, financial statements or accounts, and any sensitive inform

						domain	on Individual religious beliefs Trade union An individual sexual life/sexual preferences Racial and ethnicity Genetic data Online biometric data such as fingerprints and pictures Health data.” [49]		ation in business plans.” [50]
Color	White	Grey	Green	Red	Red	Red	Black	Black	Black
Protection	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
User Groups	No	No	No	All employees	All employees	All employees and external parties	**Internal**: from within the company to	**Internal**: from within the company to within the company **External**: from within the	**Internal**: from within the company to within

						in the list below:	within the company **External**: from within the company to outside the company	company to outside the company	the company **External**: from within the company to outside the company
permission	No	No	No	All employees are Co-Author	Custom permission of user preference	All employees are Co-Author Custom permission for the external Parties by the user	Custom permission	Custom permission	Custom permission
Offline Access	No	No	No	No	No	No	No	No	No
Visual marking	No	No	Footer: Internal Font: 8 Alignment: Left	Footer: Confidential Font: 8 Alignment: Left	Footer: Confidential Font: 8 Alignment: Left	Footer: Confidential Font: 8 Alignment: Left	Footer: Sensitive Font: 8 Alignment: Left	Footer: Sensitive Font: 8 Alignment: Left	Footer: Sensitive Font: 8 Alignment: Left
Condition	No	No	No	No	No	No	Yes	Yes	Yes