2022 International Conference on Futuristic Technologies (INCOFT)
Karnataka, India. Nov 25-27, 2022

Data Leakage Prevention System for Internal Security

Bhavya Singh Shishodia
Defence Institute of Advanced Technology,
Pune , India
bsshishodia21@gmail.com

Manisha J. Nene
Defence Institute of Advanced Technology,
Pune , India
mjnene@diat.ac.in

2022 International Conference on Futuristic Technologies (INCOFT) | 978-1-6654-5046-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/INCOFT55651.2022.10094509

***Abstract* - Transferring both allowed and illegitimate information is increasingly routine. This increased the potential danger to sensitive information and opened the door to further threats. A data breach is becoming commonplace in the headlines. All kinds of harm may be done with stolen information. A Data Leakage Prevention System (DLPS) is a scheme for preventing the unauthorised release of sensitive information inside an organization's internal network. The purpose of this study is to investigate different strategies for data security and the effects of preventing data leaks. Objective notes were taken on installation procedures and issues encountered. The deployment of industrial Data Leakage Prevention solutions in major organisations to safeguard cyber data has also been highlighted. The study holds the potential to guide a way towards implementation of technical solutions to handle the challenges envisaged in the ever-evolving environment, benefiting both academics and professionals.**

Keywords – Security, Data, Users, Information, Leakage, DLP.

I. INTRODUCTION

Only traditional firewalls will not be sufficient to ensure cyber protection. The administration of cyber security has to be approached from all angles. Of course, in order to ensure the security of our data, one need to start with the conventional security measures available in institutions and then add some creative solutions. Comparing cyberspace to a black hole is apt. This void has to be filled as much as possible using unconventional approaches. One example of these atypical answers is data leakage prevention (DLP) system.

In the field of cyber security, data-centric security—as opposed to only relying on perimeter controls—has become a primary focus for many organisations. A company's most valuable asset is its information. No matter how big or far-flung, data is the lifeblood of today's globalised businesses. The availability of highly adaptive and user-friendly data in a cost-effective package is the business case for cloud computing and related technologies. The time has passed for discussing divisions between businesses and boundaries.

In today's globalised society, information is constantly being shared from one person to the next. The first party or distributor is responsible for ensuring the integrity of the data being sent. Regardless of one's profession or age, one's data may be personally sensitive. Employees, competitors, and others may leak sensitive information for a variety of reasons. Web services, emails, cloud storage, optical discs, notebooks, and even laptop computers may all be abused by these people. Experts in data security need to devise a fool proof plan to stop this harmful data leakage. A system of data security is essential to prevent information disclosure.

A. Consequences of Data Leakage

Cybersecurity specialists have a difficult challenge of preventing data leaks without disrupting legitimate company processes.

The importance of Data Leakage Prevention (DLP) to information security initiatives is sometimes underestimated because of the time and skill required to implement it. Despite the importance of preventing data leakage, most organisations continue to fail badly at this. Facebook, Google, and many other large corporations are among those that will lose out. Most current age scams in the sector may be traced back to unauthorised data leaks [4].

B. Data Leakage Prevention Solutions

Incorporating the suggested model and technology will aid firms in spotting sensitive data and conducting in-depth investigations into document metadata in transit to ensure that no such data is being lacked. Application and email monitoring, virus prevention, and user permission are only some of the features of contemporary DLP system [5].

Data-in-transit layer security is ensured by a network-based DLP solution. Data loss prevention (DLP) network solutions are often placed at network edges to keep tabs on information as it enters and leaves an organization's internal network. Email, instant messaging, secure socket layer (SSL), and other network traffic are all monitored. First, these systems must be set up in accordance with a stated information disclosure policy that serves to separate sensitive information from other types of data. The storage media where data is stored is monitored by this form of DLP system. It functions by verifying whether or not data storage on the servers in question is in line with the company's safe-keeping standards. When a policy is broken, an alert is sent to the administrator.

Data leakage can occur in a number of ways, and an endpoint DLP solution keeps tabs on user devices to prevent this from happening. "Endpoint DLP systems may be set up to actively prevent unauthorised user behaviours and provide an instant warning to the administrator for any policy breach."

C. Related Work

Cyber-attacks have become more of a concern with older technologies [6] due to the ineffectiveness of the current security infrastructure and mechanisms. In the past, the goals of cyberattacks were limited to stealing sensitive data from users' computers or damaging their infrastructure. Recent hacking assaults, however, have shifted their focus from

978-1-6654-5046-1/22/$31.00 ©2022 IEEE

leaking information and destroying services to targeting massive systems like vital infrastructures and government institutions. These days, a company can't function without using internet technology. By transmitting information from one area to another, these businesses may boost productivity. However, there are a lot of risks associated with delivering sensitive company information since a rogue worker might potentially leak the information. The term ‘data leakage problem’ is used to describe this issue. Presenting a methodology for identifying and stopping data leaks in this work. The purpose of this approach is to track out the person responsible for leaking sensitive company information. Data allocation algorithms may be used to track out the leaking agent and pinpoint data leaks in distributed data. Increase the likelihood of identifying the guilty party and ensure the safety of the relevant information.

Information is a company's most important asset, thus protecting it is a top priority. Due to a lack of available computing resources, many businesses are turning to the cloud to take advantage of the vast amounts of processing power, data storage capacity, and even application-specific software that can be made available to employees on a pay-as-you-go basis [7]. There are many advantages to cloud computing, but users' privacy must be protected. Intentional or accidental disclosure of private data to an untrusted third party is known as ‘data leakage.’ All types of outward and inbound communication, such as email, IM, online forms, and file transfers, increase the risk of sensitive information being compromised. A data leakage prevention system (DLPS) is a method for preventing the unauthorised release of sensitive information inside an organization's internal network. This essay's goal is to assess the issue of preventing data leakage while considering the numerous challenges, data security measures already in use, as well as their benefits and drawbacks.

Data leakage prevention (DLP) [8] solutions are implemented out of the box to monitor and manage data access and consumption across storage systems, client endpoints, and networks to help with this problem. Products from industry heavyweights like McAfee, Symantec, and Websense have all developed into content-aware DLP solutions for enterprises in recent years. This study, however, contends that standard, off-the-shelf methods are not sufficient for safeguarding data. If organisational and technological criteria are met prior to deploying a DLP system, then the likelihood of a wide range of events may be reduced. This paper presents the information leakage prevention (ILP) pyramid as a framework within which DLP may be effectively implemented. In addition, data should not be merged with other types of information that has distinct security needs.

II. DATA PROTECTION FOR VARIOUS DATA STATES

Table 1 shows the different data states and the level of protection they get. The steps taken by DLPS at different points in time to ensure data security are shown in figure 1.

A. *Precautions for Handling Data in Motion*

Data at Rest: Preventing information loss by leaking data requires content discovery tools. Scan the laptop, FTP server, SMTP server, and database to find out where the sensitive data is stored [9]. The following methods may be used for content discovery:

TABLE I. DATA LEAKAGE PROTECTION FOR DIFFERENT DATA STATES

Type	Description	DLP goal
Data-at-rest	Information stored in an organization like files, servers, document management systems and email servers.	Content discovery
Data-in-motion	Organization data is restricted to network traffic such as web traffic.	Block transmission of sensitive data.
Data-in-use	Information currently used at the end points such as http, https, print, file to USB and outlooks.	Prevents unauthorized usage of data (e.g., copying to a thumb drive).

Analysing data on a local level with this method, a special agent is set up on the host system to routinely check the data contained inside the files. When it detects harmful code, it moves, encrypts, and isolates the material. During this process, agents are constantly working to implement a policy, even if the devices involved aren't physically present or connected to the network. The agents' limited processing power and little memory are disadvantages of the target system. Scanning is conducted from distant computers by keeping a connection via server and application-level protocols. "Disadvantage: higher network traffic and poor performance result from remote scanning."

Data-in-transit safeguards: The company's main hub is equipped with network-based technologies. When malicious activity is detected, the gateway computer will quickly stop it. These tools do the content analysis in real time on the whole dataset.

Safety measures for Data-in-Use: Local agents and host machines regularly check sensitive data such as data copied from one location and pasted into another location, data from print screen, unauthorized data transmission and copying data to a USB/CD/DVD [10].

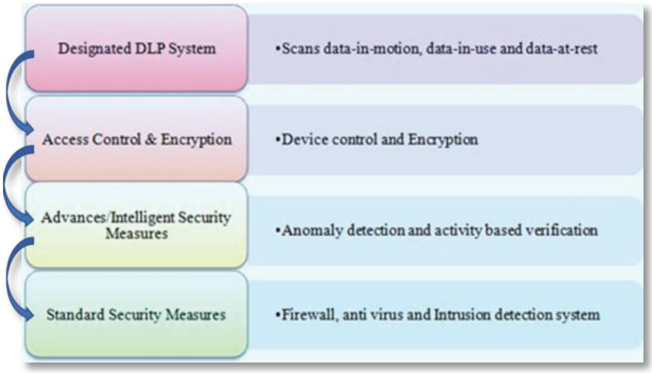

Designated DLP System
•Scans data-in-motion, data-in-use and data-at-rest
Access Control & Encryption
•Device control and Encryption
Advances/Intelligent Security Measures
•Anomaly detection and activity based verification
Standard Security Measures
•Firewall, anti virus and Intrusion detection system

Fig. 1. DLPS Activities

III. DLPS TECHNIQUES

Several studies present various DLPS paired with different technologies and methodologies in an effort to assure the best possible protection of sensitive data. The technology and DLPS methods pertinent to this subject have been explained in general terms.

A. *Intelligent documents*

This method involves encapsulating the document's data as well as the security controls that govern how that data is used[11]. The security mechanisms can be content reading, editing, or deletion, or they can need a specific authorised user for each action. With the help of this method, it is feasible to keep track of who, when, and where a document's content is accessed. It is a method that is frequently employed in DRM systems and is highly beneficial when combined with DLPS.

B. *Encryption*

Because it is the primary foundation of security and is based on the translation of data from a readable format to an encrypted format, cryptography is the approach in DLPS that is employed the most frequently. Any encryption technique is equivalent to a mutating substitution algorithm, where the substitution table is anything non-fixed and the substitution unit is the concept of "block" (and therefore mutating). The mutability of the method, which shields it from statistical attacks, contributes to its robustness [12].

C. *Hash*

A common DLPS technique is accurate file hash matching. This technique compares the hash values of intercepted communication and previously stored sensitive material to verify outbound traffic. The system detects a leak if there is a match between the values. This method has the drawback of making it impossible for the system to identify the confidential document because any changes to the original document could produce an entirely new hash value.

D. *Virtual file system (VFS)*

A virtual file system (VSF) sits atop a real file system (RFS), acting as a transitional layer between system calls and the RFS driver[13]. Additionally, they enable actions to be taken both before and after reading, writing, etc. Some of the initial RFS performance is lost in exchange for this intermediary "translation" between the applications and the file system.

E. *Minifilters*

Minifilters are low-level programmes that operate in Windows kernel mode and carry out value-added activities on filesystem operations (backup, encryption, monitoring, etc.) [14].

F. *Biometric information*

This technique is widely used in DLPS to identify the user accessing the information and thus try to ensure that it is a legitimate user with permissions to access the information [15].

G. *Hypervisor*

The method, which circumvents the dependence on pre-existing security perimeters, uses hypervisor-based memory introspection to search for the presence of sensitive raw data in memory on both client and server workstations. This technique has a high computational cost since using a hypervisor-based tool to monitor system calls necessitates the deployment of one or more virtual machines, which use an excessive amount of memory and computing power [16].

IV. METHODOLOGY

The selection criteria of DLP systems, topology drawings linked to the installation, installation phases, integration, and performance concerns have all been reviewed in depth via a genuine application since there is so little information regarding industrial DLP systems in the literature. This study project implemented a DLP system at the SSA. Following an overview of the Social Security Administration, the paper will go on to discuss DLP solution methods and selection criteria, the many phases of architecture and installation, and finally, metrics for measuring the effectiveness of the system's improved performance. The findings and the lessons gained will be reviewed once the difficulties and limitations encountered have been described[17].

A. *Monitoring and Prevention*

Monitoring feature solutions need to keep tabs on every data access in real time to spot suspicious behaviour. A large number of suppliers call for DLP solution and MTA integration if you want to use the preventive function, which allows you to monitor and block e-mails containing critical company information (Mail Transfer Agent)

B. *Centralized Management*

The loss of skilled workers is affecting almost every industry. If you're releasing a brand-new product, you'll inevitably have to buy brand-new components. Management consolidation may lessen the load on each employee. In order to save money on labour, all of your policies, reports, and data filters should be managed from a single interface.

C. *Backup and Storage Requirements*

Almost every association need data storage. Some of the DLP vendors are software-based – some of them have hardware appliances.

D. *Ease of Integration*

Each association has its special needs. Therefore, generally, box solutions of vendors do not meet the specific requirements of customers.

E. *Staffing Needs*

As was previously said, practically all businesses are understaffed. You will require more people and they will need extensive knowledge with DLP if you choose with a DLP provider that lacks a centralised administration screen, a user-friendly admin panel, or mature processes.

V. OBSERVATIONS AND RESULTS

Result of Vendor Choosing

Features were given a score between 0 and 4 based on the aforementioned concerns and the features listed in table 1. (4 for best 0 for worst). Institution selected Symantec's DLP product out of popular firms after a thorough evaluation[17].

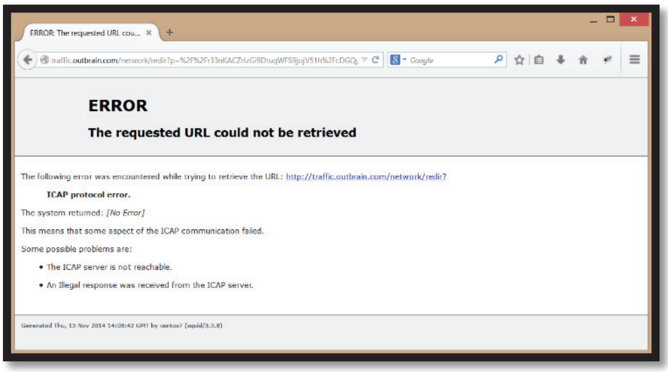
As our goal is not to choose the greatest product but to provide several DLP procedures and identify the best solution by comparing their efficacy, not engaging in a comparison like the SC Magazine that Alneyadi et al. mentioned in their study [18]. This is due to the fact that these analyses are based on case studies without a comprehensive description of attack vectors or a well-defined threat model.

A. *ICAP Integration Problem*

A flaw noticed in the DLP/web gateway integration after integrating ICAP. During peak periods, users reported seeing the message seen in Fig 2 below the screen.

TABLE II. COMPARISON OF VARIOUS FIRMS

	Symantec	Firm B	Firm C	Firm D
Product Capabilities				
Network Monitoring	4	3	3	0
Email and web prevention	4	4	4	2
Data discovery and protection	4	4	4	2
File access and usage	4	2	0	2
Endpoint monitoring and protection	4	2	4	0
Cloud email monitoring and prevention	2	2	3	0
Mobile device monitoring and prevention	4	0	2	0
Scan target coverage	4	3	2	2
Self-service remediation	3	1	0	0
Performance and scalability	4	2	4	4
Extensibility (integrations & APIs)	2	0	3	0
Policy Enforcement				
Unified policy management	3	3	3	2
Content-aware detection	4	3	3	2
Incident response workflow	4	2	3	2
Role-based access control	4	1	4	0
Reporting and analytics	4	3	2	0
Management and Security				
User authentication and identity resolution	3	0	3	4
System management and security	4	4	3	2
Market Leadership				
Customer support and success	4	1	2	2
Deployment methodology	4	0	0	1
Research and development	4	1	2	1


ERROR: The requested URL cou...
traffic.outbrain.com/network/redir?p=%2F%2Fr33nKACZrlzGi9DtuqWFS9jujV51h%2FcDGQ
Google

ERROR

The requested URL could not be retrieved

The following error was encountered while trying to retrieve the URL: http://traffic.outbrain.com/network/redir?

ICAP protocol error.

The system returned: *[No Error]*

This means that some aspect of the ICAP communication failed.

Some possible problems are:

- The ICAP server is not reachable.
- An Illegal response was received from the ICAP server.

Generated Thu, 13 Nov 2014 14:08:42 GMT by centos7 (squid/3.3.8)

Fig. 2. ICAP Integration- Error

To solve this situation, increased request limits in Network Prevent Module for Web Servers.

- Maximum Number of Requests: from 64 to 86016.
- Maximum Number of Responses: from 64 to 86016.

B. System Requirements for Agents

Data loss protection agents need at least 30 MB of RAM. Additionally, 80MB of space is needed for initial setup. While doing content detection or communicating with the Endpoint Prevent server, the DLP Agent programme briefly uses more memory. When these processes are done, memory consumption goes back to its original low levels. Until the DLP Agent transmits incident data to the Endpoint Prevent server, the data must be temporarily stored on the endpoint computer, necessitating more disc space. The DLP Agent will continue to use more disc space if the endpoint computer is unable to communicate with the Endpoint Prevent server for a prolonged length of time. Requirements and suggestions for systems 22 Endpoint computers needed for DLP Agent space when new incidents are produced. Only once the agent programme reconnects to the Endpoint Prevent server and transmits the recorded occurrences [17] is the cleared disc space made available.

C. Policy Tuning

DLP systems keep a model of either permitted (whitelisting) or prohibited (blacklisting) activity in order to differentiate between legitimate and malicious transactions. The model may be learnt from prior transactions or provided by an expert's knowledge [19].

Shockingly low incidence rates were discovered, but that was to be anticipated. Fig.3 and Fig 4 show a graph of DLP Network and Endpoint incident counts over time, broken down by policy modifications and informative actions[17].

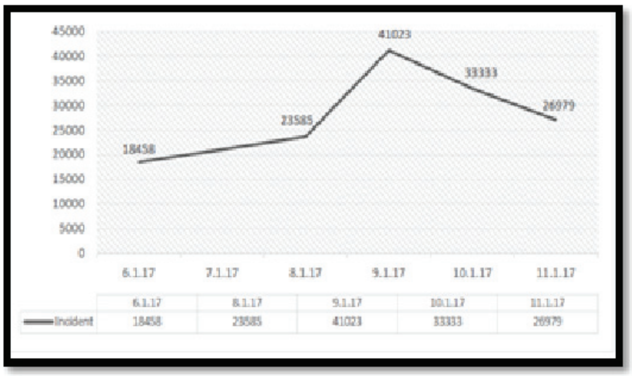


Fig. 3. DLP Endpoint Incident-Time Graphic

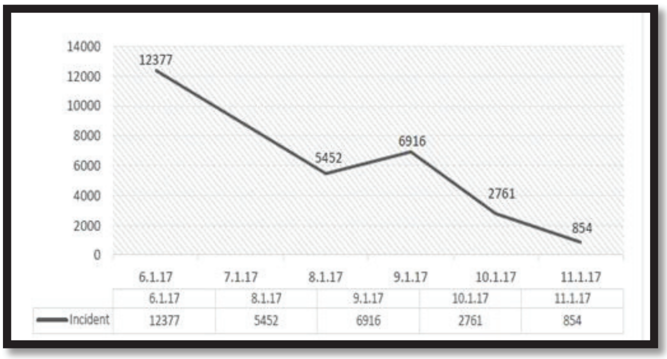


Fig. 4. DLP network incident-time graphic

Attention should be paid to the number of new clients when examining incident counts in the graphics because implementation distributed step by step. A remarkable decrease can be observed.

The studies listed below, which were partial in the study, were proposed[17].

- Implement a DLP Network monitor to look for data in motion across all protocols.
- Implement DLP Network Discover to look for data at rest.
- Implement DLP Endpoint Discover to look for data at rest in the endpoints.
- Updating the risk scores for every risky action made by users and ensuring that the DLP rules are tightened flexibly for that user according to this score.

In the first stage, the product provides regularly scheduled reports regarding software services, events, and system warnings/errors in an effort to arrive at a healthy and lasting solution. The IT department, in particular those responsible for the DLP product, reviews incidents, identifies false positives, and publishes reports after fixing identified problems. "After the first suspects have been ruled out, the IT staff and the CIO will convene to conduct a more thorough investigation." Members of the IT staff, the Chairman of IT, and the head of the executive board get together once a week to dig further into reported issues. After the system is set up, risk managers and their teams must determine whether to report the occurrence to legal authorities or tell the end user of his carelessness. The avoidance of data loss is a crucial measure for any business serious about securing its most private files. Projects aimed at preventing data loss are time-consuming and fraught with challenges. Some of the biggest challenges projects encountered were dealing with the reactions of end-users, breaking management, and inspectors' reluctance to DLP agents.

A reliable central administration panel is a must for DLP systems. Almost all businesses today are hampered by a dearth of qualified IT workers. Introducing a new product necessitates sourcing new components. The burden on employees might be lightened with a centralised management structure. In order to save money on labour, one should be able to make changes to or completely deactivate data filters, reports, and policies all from the same interface.

DLP projects, like many others, have backup and storage needs. DLP hardware or software must be able to handle terabytes of data if your data retention policy mandates data retention for many years. It's not only inefficient; it ends up costing a lot of money, too. An effective DLP system will also include a strategy for archiving and backing up data.

It is recommended to initially opt to conduct the data loss prevention project in monitor mode without installing an agent for end-users, as this will provide the healthiest outcomes and cause the least amount of disruption to the user. "You may elect to deploy agents to end users after you have command of the network, the web, and email (data in transit)." In order to effectively roll out agents, it is best to begin with the pilot units or areas. Agent-based endpoint visibility defines the exceptional by providing very granular insights into worker behaviour. You are able to detect shifts in activity patterns and properly anticipate when employees may be leaving. (If a worker is copying everything, of course).

VI. CONCLUSION

The market for data loss prevention solutions is complex, but one can choose the right tool for the needs by learning about the features offered by DLP solutions and following a systematic selection process. Incorrect assumptions and a lack of preparation for the DLP business process and workflow are the two primary technical barriers to a successful implementation.

Software for data loss prevention (DLP) can stop ineffective business processes that include sensitive data. The business is still a few years away from being able to stop well-informed bad actors, even if it can thwart certain detrimental initiatives. It is crucial to have a clear understanding of which departments will be involved and how one intends to handle infractions before beginning the selection process. One doesn't want to learn about a new acquisition's shortcomings after deployment, when the incorrect individuals are observing policy violations, if it isn't capable of protecting the sensitive data of a business unit that wasn't involved in the selection process.

Last but not least, this study aims to shed light on the deployment of industrial DLP solutions—which have gotten little attention in the existing literature—within significant organisations, as well as the prospective results of doing so. The problems encountered and their remedies are discussed.

REFERENCES

[1] Ghouse, Mohammed, Manisha J. Nene, and C. Vembuselvi. "Data Leakage Prevention for Data in Transit using Artificial Intelligence and Encryption Techniques." 2019 International Conference on Advances in Computing, Communication and Control (ICAC3). IEEE, 2019.

[2] Devi, Girija, and Manisha J. Nene. "Security breach and forensics in intelligent systems." Information and Communication Technology for Intelligent Systems. Springer, Singapore, 2019. 349-360.

[3] Perwej, Dr. Yusuf & Abbas, Qamar & Dixit, Jai & Akhtar, Nikhat & Jaiswal, Anurag. (2021). A Systematic Literature Review on the Cyber Security. International Journal of Scientific Research and Management. Volume 9. Pages 669 - 710. 10.18535/ijsrm/v9i12.ec04.

[4] Jadhav, Prasad & Chawan, Pramila. (2019). Data Leak Prevention System: A Survey. 06. 04.

[5] Hauer, Barbara. (2015). Data and Information Leakage Prevention Within the Scope of Information Security. IEEE Access. 3. 1-1. 10.1109/ACCESS.2015.2506185.

[6] Patil, C., Nalawade, S., Natekar, V.D., & Saxena, P.N. (2017). Data Leakage Detection and Prevention System.

[7] Gupta, Ishu, A Comparative Study of the Approach Provided for Preventing the Data Leakage (2017). International Journal of Network Security & Its Applications (IJNSA) Vol.9, No.5, September 2017.

[8] Hauer, B. (2014). Data Leakage Prevention - A Position to State-of-the-Art Capabilities and Remaining Risk. ICEIS.

[9] R. Tahboub and Y. Saleh, ‘Data Leakage/Loss Prevention Systems (DLP),’ 2014 World Congress on Computer Applications and Information Systems (WCCAIS), Hammamet, 2014, pp. 1-6.

[10] D. Kolevski and K. Michael, ‘Cloud computing data breaches a socio-technical review of literature,’ 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), Noida, 2015, pp. 1486-1495.

[11] Husham Ali, B., Jalal, A.A., Al-Obaydy Al-Obaydy, W.N.I.: Data loss prevention (DLP) by using MRSH-v2 algorithm. Int. J. Electr. Comput. Eng. (IJECE). 10, 3615 (2020).

[12] Holgado, P., García, A., García, J.J., Roncero, J., Villagrá, V.A., Jalain, H.: Context-based Encryption Applied to Data Leakage Prevention Solutions. In: Proceedings of the 14th International Joint

Conference on e-Business and Telecommunications. pp. 566–571. SCITEPRESS - Science and Technology Publications (2017).

[13] Hu, C., Chen, F., Zheng, H.: Researches on the Security Protection and Inspection Method for Confidential Documents Based on Linux Operating System. In: Proceedings of the 3rd International Conference on Machine Learning and Soft Computing - ICMLSC 2019. pp. 249–252. ACM Press, New York, New York, USA (2019).

[14] Buda, A., Colesa, A.: File System Minifilter Based Data Leakage Prevention System. In: 2018 17th RoEduNet Conference: Networking in Education and Research (RoEduNet). pp. 1–6. IEEE (2018)

[15] Alruban, A., Clarke, N., Li, F., Furnell, S.: Biometrically Linking Document Leakage to the Individuals Responsible. In: Furnell S., Mouratidis H., Pernul G. (eds) Trust, Privacy and Security in Digital Business. pp. 135–149 (2018).

[16] Vojnak, D.T., Eordevic, B.S., Timcenko, V.V., Strbac, S.M.: Performance Comparison of the type-2 hypervisor VirtualBox and VMWare Workstation. In: 2019 27th Telecommunications Forum (TELFOR). pp. 1–4. IEEE (2019).

[17] Symantec™ Data Loss Prevention System Requirements and Compatibility Guide – v14.6(2017).

[18] Sultan Alneyadi, Elankayer Sithirasenan, Vallipuram Muthukkumarasamy, A survey on data leakage prevention systems, Journal of Network and Computer Applications, Volume 62, 2016, Pages 137-152, ISSN 1084-8045.

[19] E. Costante, D. Fauri, S. Etalle, J. den Hartog and N. Zannone, ‘A Hybrid Framework for Data Loss Prevention and Detection,’ 2016 IEEE Security and Privacy Workshops (SPW), San Jose, CA, USA, 2016, pp. 324-333.

[20] Y. Liu, C. Corbett, K. Chiang, R. Archibald, B. Mukherjee, and D. Ghosal, ‘Detecting sensitive data exfiltration by an insider attack,’ in Proceedings of the 4th annual workshop on Cyber security and information intelligence research: developing strategies to meet the cyber security and information intelligence challenges ahead, pp. 16, 2008.