DATA LOSS PREVENTION TECHNIQUES

Dr. Uma Pujeri[1], Dr. Shamla Mantri[2], Dr. Himangi Pande[3], Priyanshu Gupta[4]

Article History: **Received:** 09.04.2023 **Revised:** 22.05.2023 **Accepted:** 28.06.2023

Abstract

The security threat of data leakage is increasing, and perimeter protection mechanisms such as firewalls are no longer sufficient to safeguard sensitive information. Data loss prevention (DLP) solutions are promising, but the lack of transparency in this area makes selecting the right solution challenging. This paper provides a systematic evaluation of content-based DLP solutions to determine their effectiveness in preventing data leakage in web traffic. The study focuses on HTTP, the dominant internet protocol, and reveals that while DLP solutions protect against accidental data loss, attackers can still bypass them. To address this challenge, The research provides an artificial neural network-based content classification technique based on MLP architecture and an n-gram TF-IDF feature descriptor to detect and protect sensitive information of a well-known TI business. When compared to existing solutions, the proposed technique is found to be significantly superior at minimizing insider threats and preventing data leaks.

Keywords: Data Loss Prevention, HTTP, Web Traffic

[1,2,3]Associate Professor, School of Computer Engineering Technology, Dr. Vishwanath Karad MIT World Peace University, Kothrud, Pune, Maharashtra, India, 411038
[4]Third Year BTech Students, School of Computer Engineering Technology, Dr. Vishwanath Karad MIT World Peace University, Kothrud, Pune, Maharashtra, India, 411038

Email: [1]uma.pujeri@mitwpu.edu.in, [2]shamla.mantri@mitwpu.edu.in, [3]himangi.pande@mitwpu.edu.in, [4]1032200918@mitwpu.edu.in

DOI: 10.31838/ecb/2023.12.s3.563

1. Introduction

Data loss prevention (DLP) is a crucial part of information security that aids organizations in defending sensitive data against theft, loss, and unauthorized access. In the current digital era, businesses produce and maintain enormous volumes of sensitive data, such as financial information, intellectual property, and customer information. Modern organizations depend on this data to function, and its loss or theft may have serious repercussions, including monetary loss, reputational harm, and legal responsibility. Therefore, for any organization that wishes to safeguard its sensitive information and guarantee business continuity, putting efficient data loss prevention procedures into place is essential.

Data loss prevention is the use of a combination of technologies, procedures, and policies to guard against unauthorized access to, use of, or disclosure of sensitive data. These precautions often combine administrative controls like rules, processes, and training programs with technological controls like encryption, access controls, and data monitoring.

Finding sensitive data within an organization is one of the main objectives of data loss prevention. This entails making a complete inventory of all available data sources and determining the categories of information that are most sensitive and important to the organization. This might be monetary information, commercial secrets, or private data like credit card numbers or social security numbers.

Once sensitive data has been located, the following step is to categorize it according to how sensitive or risky it is. In order to do this, data must be given labels or tags that specify its level of availability, secrecy, and integrity. Data could be labeled as "highly confidential" for instance if it includes trade secrets or other sensitive information that, if stolen or leaked, might have a major effect on the organization.

Following the identification and classification of sensitive data, the next step is to put in place the necessary safeguards to prevent unauthorized access or disclosure. This might be administrative controls like rules and procedures that restrict how sensitive data is accessed, utilized, and shared or technological controls like encryption or access controls. For instance, a company may put in place stringent access controls that only let staff with valid business needs to access sensitive information.

Continuous monitoring and upkeep of security controls is a crucial part of preventing data loss. To guarantee that security measures are operating correctly and that any vulnerabilities or weaknesses are quickly fixed, this calls for routinely reviewing and testing security controls. Organizations may detect security risks and take steps to fix them before they result in data breaches or other security events by conducting regular security audits, vulnerability scanning, and penetration testing.

And last, an essential part of any successful data loss prevention program is training staff members on data protection best practices and policies. Regular security awareness training, phishing simulations, or other staff education measures might be used to achieve this. Employees must comprehend the significance of data security and privacy and be given the information and skills necessary to do so effectively since they play a crucial role in securing sensitive data.

In conclusion, any comprehensive information security program must include data loss prevention. Organizations may safeguard their company operations, their clients, and their reputation by identifying and guarding sensitive data against unauthorized access, theft, or loss.

1.1 Paper Organisation

The contents of the paper are organised as follows: The significance of this topic is discussed in section II. Section III discusses the literature survey. Section IV discusses the recommended technique. Section V delves into the specifics of design and technology. Section VI discusses the benefits, whereas Section VII discusses the drawbacks. Section VIII goes over the applications. Section IX discusses the future scope, while Section X provides the conclusion.

2. Literature Review

The paper “A deep learning model for information loss prevention from Multiple page digital documents” proposes a deep learning model for data loss prevention from multi-page digital documents, A Deep Learning Model for Information Loss Prevention from Multi-Page Digital Documents. It employs an aggregate of convolutional and recurrent neural networks to extract and classify textual content and photo content from multi-web page documents. Results show accurate detection and classification of sensitive information, reduced manual labor and time required for document review, improved data security, and compliance with regulations and privacy laws. However, there is limited availability of large-scale annotated datasets for multi-page digital documents, making it difficult to train deep learning models effectively. Limited research on the transferability and generalization of deep learning models across different types of documents, languages, and domains, and limited exploration of the interpretability and explainability of deep learning models for information loss prevention may limit their adoption in certain applications and industries.

The paper titled "Can Content-Based Data Loss Prevention Solutions Prevent Data Leakage in Web Traffic?" uses Deep packet inspection and machine learning techniques to analyze the content of web traffic. Predefined rules or custom machine learning models are used to identify sensitive or confidential data and prevent it from leaving the network. Results include improved data security, compliance with data privacy and security regulations, and avoidance of costly data breaches. However, there is a research gap due to limited real-world effectiveness research of content-based solutions, insufficient exploration of limitations and vulnerabilities, and the need for integrated approach research with other security measures.

The research paper "Data Loss Prevention "consists of an intensive literature assessment and statistics collection from various companies, information evaluation to identify tendencies and problems, and an advice generator to offer insights into modern-day techniques for stopping records loss and recommend improvements to data safety. The results of the paper include protecting sensitive information, preventing financial losses, and improving brand reputation. However, there is a research gap on human factors in DLP, a lack of standard evaluation metrics, and a limited understanding of DLP implementation challenges.

The paper "Predicting the likelihood of legitimate data loss in email DLP" discusses the research gap in predicting legitimate data loss in email. It involves collecting email data that contains valid and potentially risky content and developing a machine-learning model to predict the likelihood of valid data loss. The results of the research show that data protection can be improved without having to implement strict policies by recognizing legitimate data loss. However, there is a lack of research on predicting legitimate data loss in email, limited datasets for training and evaluating machine learning models, and lack of standardized evaluation metrics and methodologies for comparing different models

The paper "Data Exfiltration Techniques and Data Loss Prevention System" focuses on analyzing different data exfiltration techniques used by attackers, reviewing existing data loss prevention (DLP) systems, developing a comprehensive DLP framework that addresses common exfiltration techniques, implementing and testing the framework using various attack scenarios, and evaluating its effectiveness in preventing data loss. The research gap is limited research on exfiltration techniques used in combination to bypass DLP systems, lack of standardized evaluation methods for comparing DLP systems, and limited understanding of how attackers exploit vulnerabilities in DLP systems.

The research paper" Data loss prevention by using MRSH-v2 algorithm" focuses on data loss prevention by using the MRSH-v2 algorithm. It involves gathering data samples from various sources, preprocessing the data using feature extraction methods, training a machine learning algorithm (MRSH-v2) on the preprocessed data, evaluating the performance of the algorithm, and contrasting outcomes with other cutting-edge data loss prevention methods. The results show that MRSH-v2 is a highly accurate algorithm for detecting potential data loss incidents, is efficient and can process large volumes of data in real-time, and can help organizations prevent data loss incidents and protect sensitive information. However, there is limited research on the use of the MRSH-v2 algorithm for data loss prevention, and the effectiveness of the algorithm in detecting new and emerging data exfiltration techniques is unknown. There is a need to validate the performance of the algorithm across different types of datasets and scenarios.

The paper "Context-Aware Data Loss Prevention for Cloud Storage Services" focuses on context-aware Data Loss Prevention for Cloud Storage Services. It defines data context and designs context-aware rules, implements the rule engine, tests cloud storage data and evaluates system performance. The result is that it provides a high level of security by identifying sensitive data in the cloud, enables the implementation of granular access control policies for cloud data, and improves compliance with data protection regulations. However, the research gap is that the solution was not evaluated against various attack scenarios and takes into little account of user behavior.

3. Methodology

Data loss prevention (DLP) is a process that includes a number of policies, practices, and technology to prevent unauthorized access to, use of, or disclosure of sensitive data. Identification of sensitive data, classification, the implementation of measures to safeguard it, and ongoing monitoring of the controls to assure their efficacy are often included in the approach for DLP.

Any DLP process starts with identifying the data that needs to be protected. This entails identifying any sensitive information, including financial, healthcare, intellectual property, or data containing personally identifiable information (PII). Tools for data discovery and categorization are used to search through the data repositories of an organization to find sensitive material depending on its location, nature, and degree of sensitivity.

Once sensitive data has been identified, it is then classified based on its level of sensitivity. Data can be

labeled based on sensitivity levels or categorized based on risk levels. Data classification helps organizations prioritize their efforts to protect the most sensitive data and ensure that appropriate controls are put in place.

Following that, organizations put in place safeguards to secure sensitive data. This includes deploying encryption techniques to safeguard data at rest and in transit, access controls to limit sensitive data access to authorized individuals only, and endpoint protection measures to secure the devices used to access sensitive data.. Cloud security techniques can also be used to secure data stored in the cloud.

Continuous monitoring is essential to guarantee the efficacy of controls after they are put in place. To identify unusual behavior that could be a sign of data loss or exfiltration, regular vulnerability scans, log analysis, and other monitoring techniques are utilized. In order to pinpoint areas that need improvement and make sure that the controls are kept up to date throughout time, regular security evaluations are also carried out.

Last but not least, it is critical to inform staff members of the value of data protection and their part in preventing data loss. This involves instructing the best ways to handle sensitive data, such as avoiding using unprotected devices or networks, creating strong passwords, and alerting IT or security staff to any questionable activities.

The identification and classification of sensitive data, the implementation of measures to safeguard it, ongoing monitoring and maintenance of those controls, and staff training are all crucial phases in the DLP process. Organizations may successfully safeguard their sensitive data and avoid data loss by adhering to these guidelines and utilizing a combination of technology and best practices.

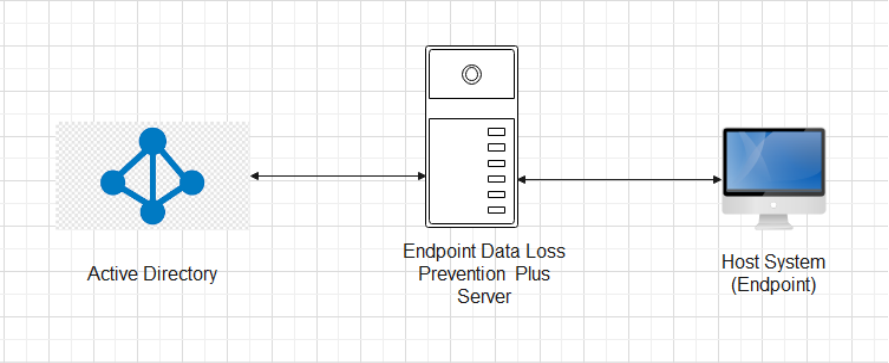

Fig 1: Endpoint DLP Plus LAN Architecture

3.1 Endpoint DLP Plus Server

The Endpoint DLP Plus server is located on the customer`s site. This server facilitates the deployment of the Endpoint DLP Plus policies defined to discover and classify data as well as determine boundaries within which the data should be secured.

PORT	PURPOSE	TYPE	CONNECTION
8090	for communication between AGENT and the endpoint DLP PLUS SERVER	HTTP	Inbound to the server
9083	for communication between AGENT and the endpoint DLP PLUS SERVER	HTTPS	Inbound to the server
6452	AGENT server communication	TCP	Inbound to the server

Fig 2: Server

3.2 Agents

The Endpoint DLP Plus agent is a software application that is installed in computers managed by Endpoint DLP Plus. It is automatically installed in LAN computers and aids in completing tasks initiated from the Endpoint DLP Plus server. For instance, if you need to blacklist or whitelist an application in a computer group, you can configure the settings in the Endpoint DLP Plus server, and the agent will ensure the task's effective completion. Moreover, the agent provides reports and audits of all the running applications on the computers, updating the Endpoint DLP Plus server. The agent contacts the server every 90 minutes for a refresh interval

3.3 Web browser

All apps operating on the controlled systems may be handled by administrators from a single location using the Endpoint DLP Plus Web console. Through the Internet or a VPN, it is reachable from

everywhere, including LAN, WAN, and distant sites like home. Additionally, there is no need for separate client installs in order to access the Web interface.

3.4 Dynamic Directory

The Endpoint DLP Plus server gathers information from Active Directory in a domain configuration that uses Active Directory to create reports on a variety of topics, including Sites, Domains, Organizational Units (OUs), Groups, and Computers. Administrators may now easily access all pertinent data kept in Active Directory for efficient domain administration.

4. Details of Technology

4.1 Identify Sensitive Data

Technology works by detecting and finding sensitive information inside an organization. It can seek text or trends that disclose sensitive information and utilize automated scanning techniques to hunt for specific information like social security numbers or credit card numbers. It is used in Data Loss Prevention to protect data against theft or unauthorized access. Advantages include adhering to data protection rules, better managing and protecting sensitive data, and lessening the chance of a data leak. Disadvantages include being expensive to use and maintain, and installation and configuration may take time.

4.2 Classify Sensitive Data

Technology such as Classify Sensitive Data is used to categorize sensitive data according to significance or sensitivity levels. It is used in Data Loss Prevention to help companies prioritize their data security efforts and ensure appropriate security controls are in place for each type of sensitive data. Advantages include making it possible to prioritize data protection operations and effectively manage risks, as well as making sure the proper security measures are in place for each sort of sensitive data. Disadvantages include creating and maintaining an efficient categorization system, as well as a challenge to establish uniform categorization across multiple departments or business divisions.

4.3 Implement Controls

Data Loss Prevention (DLP) is a technology used to protect sensitive data from unauthorized access, theft, or loss. It provides effective measures for data protection and allows organizations to customize security measures to meet specific needs and dangers. However, it can be costly to use and maintain and policies and procedures must be established and kept up with a lot of work. Advantages include providing effective measures for data protection and enabling organizations to customize security measures to meet specific needs and dangers. Disadvantages include being costly to use and maintain and policies and procedures must be established and kept up with a lot of work.

4.4 Monitor and Maintain Controls

Technology is used to monitor and maintain security controls to ensure that they are working effectively. This includes regular security audits, vulnerability scanning, or penetration testing. Data Loss Prevention is used to identify and address security vulnerabilities and weaknesses in an organization's data protection measures. Advantages include ensuring that security controls remain effective over time, enabling organizations to quickly identify and address security issues before they become major problems, and requiring specialized expertise to conduct effective security audits and testing. Disadvantages include being time-consuming and resource-intensive and requiring specialized expertise.

4.5 Educate Employees

Educating Employees is a technology that involves training and educating employees on data protection best practices and policies. It can include regular security awareness training, phishing simulations, or other forms of employee education. It helps to reduce the risk of human error and employee negligence that can lead to data breaches. However, it may be time-consuming and costly to develop and deliver effective training programs, and may not be effective if employees do not take the training seriously or understand its importance.

4.6 Deep Learning Model (MLP)

The Multilayer Perceptron (MLP) is a type of artificial neural network commonly used for various tasks such as classification and pattern recognition. Its working involves three main components: the input layer, hidden layers, and the output layer. The input layer receives the input data, which is represented as a vector of features. Each feature corresponds to a node in the input layer. The hidden layers, which can consist of one or more layers, perform computations on the input data through a series of weighted connections and activation functions. These computations enable the network to extract and learn complex patterns from the data. Finally, the output layer produces the network's predictions or outputs based on the computations performed in the hidden layers. Through an iterative training process, the MLP adjusts its weights to minimize errors and improve its ability to accurately classify or predict new data.

4.7 Multi-Resolution Similarity Hashing (MRSH-v2)

This algorithm is carried out in two phases. Stage 1 involves the creation of a fingerprint/signature by extracting specified features from previously known files. Stage 2 entails comparing the fingerprints of new files to the fingerprints acquired from the first stage. The comparison yielded results ranging from 0 and 100. To put it another way, when the probability

outcomes of comparing two files are high, the files are comparable. When these ratios are low, files are not similar to one another.

5. Results

The deep Learning model Multilayer Perceptron (MLP) was evaluated on 68,099 text documents chosen at random. The end result was as follows: Accuracy was 0.8920, F1-Score was 0.8751, Precision was 0.9281, and Recall was 0.8279. On the TS dataset, the MRSH-v2 method is tested. This dataset contains a variety of file kinds, including pdf, excel, doc, gif, xls, ppt, and txt. 290,314 packets were generated to transmit 3000 files from the TS dataset across the network. The MRSH-v2 algorithm properly detected 249670 of the 290314 packets generated, which contained various types of files in the dataset. MRSH-v2 detected 40643 packets carrying various types of files. MRSH -v2 did not detect 29031 packets containing files. As a result, the accuracy is 0.85. The data classification technique was evaluated on the MNIST dataset which gives accuracy rates of over 99%. Particularly convolutional neural networks (CNNs) have been particularly successful in accurately classifying the handwritten digits in this dataset. The MNIST dataset was used to test the data classification technique, which yielded accuracy rates of more than 99%. Convolutional neural networks (CNNs) in particular have been very successful in correctly categorizing the handwritten digits in this dataset.

6. Conclusion

Any organization's security strategy must include data loss prevention (DLP) as a key strategy. DLP assists organizations in safeguarding their sensitive data from both internal and external threats by preventing unauthorized access, loss, or theft of data. The identification, control, monitoring, and detection of data are all parts of a comprehensive DLP strategy. The best DLP tools can reduce risks, enhance compliance, and safeguard an organization's reputation. Investing in DLP can help organizations save time and money in the long run because, as with many things, prevention is always preferable to cure. DLP aids businesses in adhering to rules and regulations to stay out of trouble. DLP also guards against insider threats, which are a significant cause of data breaches. It is crucial to remember that DLP solutions need to be updated and tested frequently to remain effective because threats and technologies change over time. Organizations that prioritize DLP and keep a proactive security approach will ultimately be better able to protect their priceless data assets and keep the confidence of their stakeholders.

Future Scope

In the future, the field of data loss prevention (DLP) will witness several advancements. Deep learning algorithms, a form of machine learning, will play a crucial role in analysing data usage patterns and proactively responding to potential security threats. Cloud-based DLP solutions will offer enhanced data security and flexibility, enabling organizations to protect their sensitive information across diverse platforms and devices while easily scaling their security measures as needs evolve. The emergence of blockchain technology will revolutionize DLP by providing a decentralized and tamper-proof network, ensuring data integrity and preventing breaches. Incorporating DLP techniques into the rapidly expanding Internet of Things (IoT) ecosystem will ensure secure data transmission and safeguard against unauthorized access. With the advent of quantum computing, quantum-resistant cryptography will become a vital aspect of DLP, protecting sensitive data from the vulnerabilities posed by quantum computers. Additionally, the importance of privacy will continue to grow, leading to the integration of DLP techniques during the design phase of products and services, ensuring privacy is a fundamental consideration from the outset. These advancements in DLP hold great promise for a future where data is comprehensively protected, and privacy is prioritized in every aspect of technology.

7. References

1. A. Guha, D. Samanta, A. Banerjee and D. Agarwal, "A Deep Learning Model for Information Loss Prevention From Multi-Page Digital Documents", 2021, doi: 10.1109/ACCESS.2021.3084841.
2. D. Gugelmann, P. Studerus, V. Lenders and B. Ager, "Can Content-Based Data Loss Prevention Solutions Prevent Data Leakage in Web Traffic? July-Aug. 2015, doi: 10.1109/MSP.2015.88.
3. S. Liu and R. Kuhn, "Data Loss Prevention," in IT Professional, March-April 2020, doi: 10.1109/MITP.2010.52.
4. Mohamed Falah Faiz , Junaid Arshad , Mamoun Alazab , Andrii Shalaginov, “Predicting likelihood of legitimate data loss in email DLP”, September 2020, doi: https://doi.org/10.1016/j.future.2019.11.004
5. H. AlKilani, M. Nasereddin, A. Hadi and S. Tedmori, "Data Exfiltration Techniques and Data Loss Prevention System," 2019, doi: 10.1109/ACIT47987.2019.8991131
6. Basheer Husham Ali , Ahmed Adeeb Jalal, Wasseem N. Ibrahem Al-Obaydy, “Data loss prevention (DLP) by using MRSH-v2 algorithm” August 2020 International Journal of Electrical and Computer Engineering 10(4):3615 DOI:10.11591/ijece.v10i4.pp3615-3622
7. T. Wüchner and A. Pretschner, "Data Loss Prevention Based on Data-Driven Usage Control,"

2012 IEEE 23rd International Symposium on Software Reliability Engineering, Dallas, TX, USA, 2019, pp. 151-160, doi: 10.1109/ISSRE.2012.10.

8. Y. J. Ong, M. Qiao, R. Routray and R. Raphael, "Context-Aware Data Loss Prevention for Cloud Storage Services," 2020 IEEE 10th International Conference on Cloud Computing (CLOUD), Honolulu, HI, USA, 2020, pp. 399-406,doi: 10.1109/CLOUD.2017.58.