

Pragmatic Study of Enterprise Data Breach and its Prevention Technique by DLPD

Dr. Premasudha B G¹, Vijayalakshmi S Katti² and Srivyshnavi H M³

¹Professor and Head, Department of MCA, Siddaganga Institute of Technology, Tumkur, Karnataka, India
Email: bgpremasudha@sit.ac.in

²Research Scholar, Department of MCA, R. V. College of Engineering, Bengaluru, Karnataka, India
Email: viju.s.katti@gmail.com

³Student, Department of MCA, Siddaganga Institute of Technology, Tumkur, Karnataka, India
Email: srivyshnavi261999@gmail.com

Abstract—Data Breach is a security problem in which bad guys gain access to confidential or sensitive information along with personal data like social security numbers, bank account numbers, healthcare data or corporate data which include customer data records and financial information. It can be accidental or purposely exposed to people. Nowadays, data is super important for businesses. If data leaks out, it can really be a big delinquent to a company's reputation and a big financial loss. With more and more data being created, and data breaches happening a lot, stopping data leaks is a big challenge. This review helps readers to learn about data leaks, threats, recent data leaks and various prevention techniques.

Index Terms— Data Breach, sensitive information, Data threats, DLPD techniques.

I. INTRODUCTION

Data leakage is a big problem for big companies, government groups and small companies, as it doesn't depend on the size of the company. When sensitive information gets out by mistake or because of bad people, it can make a company look bad and lose a lot of money. It can even make the company unstable in the long run. The stuff that gets leaked can be things like info about employees or customers, secret business ideas, or medical records. In 2016, IBM found that, on average, a data breach cost companies about \$4 million! And a prediction from Juniper Research said that because more and more of our lives and company stuff are on computers, the cost of data leaks worldwide could be over \$2.1 trillion by 2019. In 2023 it was 4.45 million and it has increased 15% over 3 years. [1]

Data leakage or loss can be caused by both internal and external information breach, either intentionally or in an unintentional way. A study from Intel showed that employees who are working there that are internal account 43% of corporate data breach, and half of these are unintentional or accidental. Motivations for insider attacks are varied, which includes corporate spying, unfair to their employer or financial reward. Accidental leaks mainly occur due to poor business processes such as failure in applying relevant technologies and security policies. [6, 9]

Data loss prevention means using different kinds of methods to keep important and sensitive information safe. In which it includes rules, technology, and training to prevent unauthorized people from getting, sharing, or using sensitive data with technological controls like encryption, access controls and data monitoring. [6]

Data Loss prevention tools and techniques are very important because they solve a unique problem, and the people may not be same to handle other security tasks. Sometimes, non-technical people like legal officers oversee protecting and safeguard the data. Even human resources can be involved in dealing with DLP alerts. Some companies find the rules for DLP need to be seen by leaders of different business units not only security teams. [7]

It is challenging for companies to secure data against data breaches in the era of big data. As data is a very important and crucial component of the enterprise, managing and securing large amounts of data is a challenging thing. The need to store, analyze and process more data together including modern communication channels results in an increase of data leakage vectors including email, web pages, database, media, camera, laptop theft and social networks.

The purpose of this review paper is to highlight data leak threats, prevention, and detection of data breach. At first discussed enterprise data leak threats, and a few recent data breach incidents, then we spoke about DLPD techniques.

II. LITERATURE REVIEW

The deliberate or unintentional disclosure of private data to unapproved parties is known as a data breach. Data has emerged as one of an enterprise's most vital components in the digital age. [1] The organizations are incurring bulk amount of money and efforts, right after the data breaches, expenses, and efforts from lawsuits, installing cyber security measures and organization policy changes.[2] In order to detect confidential data in the organization, DLP systems are typically implemented using various mechanisms, such as Described Content Matching (DCM), exact data matching, indexed document matching, vector machine learning, content matching classifiers, and file type detection. [3] DLPs use different techniques to analyse the content and the context of confidential data to detect or prevent leakage. Although DLPs are increasingly being designed and developed as standalone products by IT security vendors and researchers, the term still ambiguous.[4] Our privacy-preserving data-leak detection method supports practical data-leak detection as a service and minimizes the knowledge that a DLD provider may gain during the process. [5]

III. ENTERPRISE DATA LEAK THREATS

A. Classification of Data Threats

Data leaks can be categorized based on their causes and culprits. Intentional leaks occur when either external actors, like hackers using tricks to breach systems, or malicious insiders, such as disgruntled employees, purposely release sensitive data. By figure 1 On the other hand, accidental leaks happen due to employees' mistakes, like sharing confidential information unintentionally or not properly protecting it. These dual factors contribute to the complexity of data leak threats, involving both insiders and outsiders, with intentions ranging from financial gain to inadvertent errors in handling sensitive information. More than 60% of leaks happens because of people inside the organization did something wrong. This shows that it's not just about technology; it's also about how people behave and making sure they do the right things to stop data leaks. [1, 7, 9].

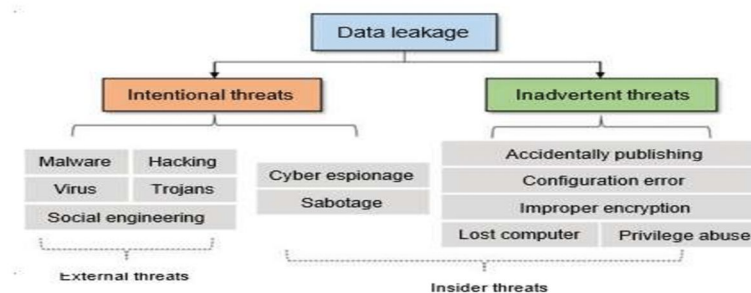


Figure1. Classification of enterprise data leak threats.[1]

Different cyber security reports might show slightly different numbers due to using different datasets, but they all agree on one trend: insider threats, meaning breaches caused by people within the company, are becoming a leading cause of data leaks, accounting for more than 40% of all breaches. [6] During the first quarter of 2023, more than six million data records were exposed worldwide through data breaches. Since the first quarter of

2020, the highest number of exposed data records was detected in the fourth quarter of 2020, nearly 125 million data sets. [8]

TABLE I. STATISTICS OF DATA LEAK INCIDENTS IN RECENT YEARS

Year	Number of incidents
2021	1,108
2020	1,001
2019	1,473
2018	1,244
2017	1,579
2016	1,093
2015	781
2014	783
2013	614
2012	470
2011	855
2010	662

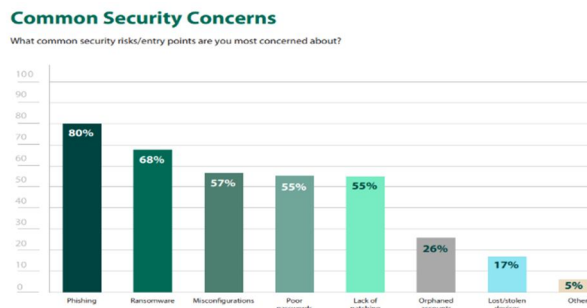


Figure2. Statistics of data leak incidents in 2023

B. Classification of Enterprise Data Leak

Internal Data Leak

An internal data leak, referred to as an insider data leak or insider data breach, happens when people working within an organization, like employees, contractors, or those with authorized access, reveal sensitive or confidential information to unauthorized parties or the public, either on purpose or by accident. This can result in security risks, financial losses, harm to an organization's reputation, and legal consequences. [8]

An employee at the Australian Red Cross Blood Service inadvertently uploaded documents containing the personal information of over 550,000 blood donors to a publicly accessible directory on their website. This sensitive data pertained to donors between 2010 and 2016 and encompassed details like names, addresses, and birthdates, along with information about their sexual history, drug usage, and medical backgrounds. Similarly, in 2011, a server in Texas unintentionally made the personal information of 3.5 million citizens available online for an entire year. [1]

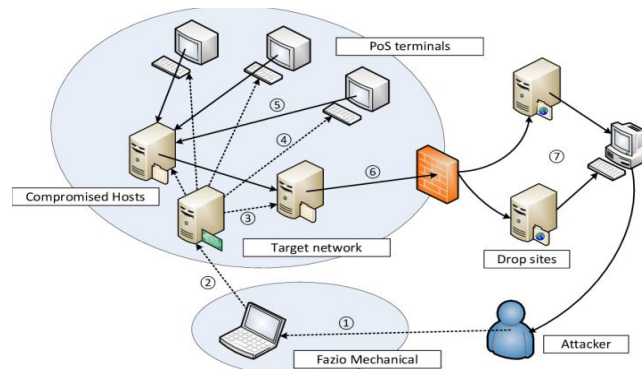


Figure 3. Breakdown and analysis of the Target Data Breach

External Data Leaks

An external data leak, referred to as an external data breach, happens when sensitive or private information is exposed or obtained by individuals or entities who shouldn't have access to it and are located outside the organization. These breaches are often the outcome of cyber attacks or security events initiated by external sources like hackers, cybercriminals, or malicious actors.

Many famous data breaches have resulted in organizations losing hundreds of millions of dollars. For example, let's look at the data breaches that occurred at Yahoo and Target. Additionally, there was a situation where the Australian Red Cross Blood Service accidentally revealed the personal details of more than 550,000 blood donors on their public website. This information contained things like names, addresses, birthdates, and even private information about donors' sexual history, drug use, and medical backgrounds. In 2011, a server in Texas also mistakenly exposed the personal information of 3.5 million citizens online for a whole year. The below listed incidents in Table 2 led to significant financial damage for the affected organizations.

TABLE II. MASSIVE ENTERPRISE DATA LEAK INCIDENTS IN RECENT YEARS

Year	Incident	Number of Records Leaked
2021	Facebook-Personal data of 533 million users	533 million records
2020	Marriott International – Guest records breach	5.2 million records
2019	Capital One – Credit card application data breach	100 million records
2018	Facebook – Cambridge Analytica scandal	87 million records
2017	Equifax – Massive credit reporting agency breach	143 million records
2016	Yahoo – Data breach of user accounts	1 billion records
2015	Anthem – Health insurance records breach	78.8 million records
2014	eBay – Personal data breach	145 million records
2013	Target – Retailer payment card data breach	40 million records
2012	LinkedIn – Password database breach	165 million records

IV. DLPD TECHNIQUES

Technological methods used in Data Loss Prevention (DLP) into two groups: one based on analyzing the content and the other based on analyzing the context. The ability of products to analyze deep content using various techniques is referred to as content-based analysis. This is distinct from analyzing context. Content-based analysis involves examining the actual content itself to prevent inadvertent data exposure in various scenarios, such as during storage, usage, or transmission. While content scanning is effective at preventing accidental data leaks, it may not be as effective when dealing with individuals, whether they are within the organization or external attackers, who attempt to hide the data. [1, 2].

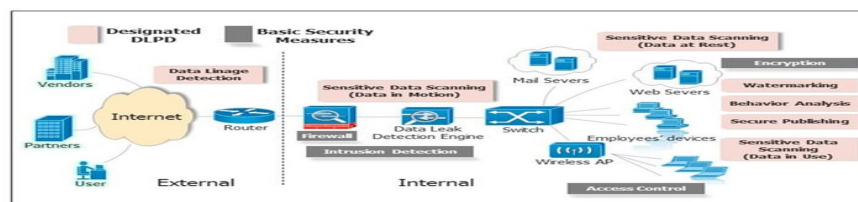


Figure 4: Depicts the multiple points for deploying complementary data leak prevention and detection (DLPD) techniques in an enterprise environment.[1]

Instead of trying to specifically detect sensitive content, context-based approaches primarily focus on analyzing the contextual information related to the monitored data or the circumstances surrounding it. Some Data Loss Prevention (DLP) solutions are a combination of both content and context analysis. Context encompasses factors such as the source, destination, size, recipients, sender, header details, metadata, time, format, and essentially anything except the actual content of the information. Contextual analysis is incredibly valuable, and any effective DLP solution should incorporate this aspect as part of its comprehensive approach. [2, 3]

A. Content based Approches

1. Snort:

Snort is like a digital security guard for computer networks, and its open source, meaning anyone can use it. It lets people create special rules to find specific patterns in network traffic, like credit card numbers or medical codes as shown in figure 4. When Snort watches the network, it checks the data passing through it against these

rules to see if anything looks suspicious, like someone trying to leak data. This way of checking data is used in many security tools, including ones that protect against data leaks. It's good at finding certain kinds of structured data, like credit card numbers or medical codes. It's fast and easy to set up because it comes with pre-made rules, but sometimes it makes mistakes and thinks something is a problem when it's not. Also, it's not great at protecting unstructured data, like important company secrets. [6, 7, 8].

Regular Expression for Credit Card Numbers

Figure5. Regular expressions for credit card numbers

2. Database Finger printing:

Finger printing is a technique that focus on finding and tracking important confidential information in a document or database. It does this by creating a unique fingerprint or signature for the crucial data while ignoring less important parts. This makes it better at detecting confidential information even if it's expressed differently. In simpler terms, it's like recognizing a specific pattern in a sea of information, helping you to find what's truly important while ignoring the rest. It is best suited for structured data from databases. [6, 7, 8, 9]

3. Exact File Matching:

This technique creates a unique fingerprint or hash value and then compares files that have the exact fingerprint, which is useful for media and binary files where the content is challenging. It also works for any file type, and it will be accurate while identifying the matching files with larger hash value when somebody modifies file slightly to avoid detection. It is effective for identifying instances when the file has sensitive content and that is duplicated or accessed by irrelevant people. [9, 10]

4. Partial Document Matching:

This technique checks if someone is trying to share the whole or partial part of document. For example, if you have done any secret business plan, the system will alert if someone tries to send a single paragraph of it to something like Instant message. Most of the solutions are based on a technique called cyclical hashing, where you take a hash bit of content, equally predetermine the number of characters, and then take another bit, and it goes through until the document is completed. This technique is best with text such as CAD files and source code and unstructured content that is sensitive. It doesn't depend on complete documents or large documents. [1, 6, 11]

5. Statistical Analysis:

Using machine learning, Bayesian analysis, and statistical techniques involves analyzing a collection of content to discover policy violations in content that looks like the protected content. This method covers a wide range of statistical approaches that can differ in how they are used and how well they work. Some of these techniques are quite like those used to filter out spam emails. This is best suited for unstructured data, but it is less effective for partial documents. It can also work for nebulous content where we may not be able to isolate the exact document. [1, 6, 12]

6. Lexical Analysis:

This technique uses a mix of tools like dictionaries, rules, and analysis to protect abstract content that looks like an "idea." For example, it can detect things like insider trading, sexual harassment, or running a private business from a work account by checking for specific words, word counts, and where those words appear in the content. It's perfect for finding violations in content that doesn't fit into neat categories or isn't found in known documents or databases. This method can catch policy violations that are loosely defined or not even considered by other monitoring techniques. [1, 6, 11]

7. Categories:

Pre-built categories with rules and dictionaries for common types of sensitive data, such as credit card numbers/PCI protection, HIPAA, etc. It is best for anything that neatly fits a provided category. Typically, it is easy to describe content related to privacy, regulations, or industry-specific guidelines. [1, 2, 10]

B. Context Based Approches

In order to address insider threats, a number of studies have analysed typical user habits. Make a model of typical data access patterns and use it to trigger alarms when anomalies arise. Detect unusual access patterns in

relational databases, with a focus on role-abusers. Create algorithms to recognize sluggish signals that point to insider threats. Incorporate anomaly- and signature-based methods to track user behaviour and identify data leaks. Captures the semantics of user intent to make sure system behaviour is consistent with user goals and stops malware from carrying out harmful operations. Keep an eye on insider activity to spot dishonest insiders using their privileges outside of their assigned role. All of these initiatives work together to help identify and reduce insider threats in organizations.

TABLE III. INFORMATION ABOUT THE DIFFERENT TECHNIQUES USED, STRENGTHS AND WEAKNESS

Technique	Description	Strengths	Weakness
Data Loss Prevention (DLP)	DLP solutions monitor and protect sensitive data to prevent unauthorized access, sharing, or leakage. They use policies and rules to detect and block data breaches.	Granular control over data – Real time monitoring and alerts, prevents intentional and accidental leaks	Complex and configure to maintain, false positives and negatives, this may not protect against zero-day threats.
Encryption	Data encryption involves converting data into a code to make it unreadable without the proper decryption key.	Strong protection for data at rest and in transit, mitigates data exposure in case of breaches.	Doesn't prevent data access, only protects data content, Key management complexity and Performance
Access control and Authentication	Access control restricts who can access specific data or resources, and authentication verifies user identities before granting access.	Strong defense against unauthorized access - Granular control over user permissions - Can prevent insider threats.	Can be bypassed if authentication credentials are compromised, May be challenging to manage permissions at scale, Doesn't address data leakage after access is granted.
Network Security	Uses firewalls, IDS/IPS, and network monitoring to protect the network perimeter and detect and block suspicious activity.	Shields the network from external threats, helps identify and block threats in real-time.	Cannot prevent all types of breaches, may not address insider threats.
User Training and Awareness	Educates employees on security best practices, phishing, and social engineering awareness to reduce human error.	Low –cost and effective in mitigating human-related vulnerabilities.	Not foolproof, relies on user compliance and awareness, this will not prevent all attacks.
Behavioral Analytics	Analyzes user and entity behavior to detect anomalies or unusual patterns of activity within the network.	Detects insider threats and advanced attacks that evade traditional security measures.	Requires historical data for accurate profiling, can generate false alarms if not tuned correctly.

By comparing all these, DLPD technique is better. Because of all these DLP offers strong capabilities for data breach prevention, organizations should evaluate their unique requirements, budget, and resources when deciding on the most suitable cyber security techniques and technologies. In practice, a combination of security measures, including DLP, encryption, access controls, and network security, is often necessary for comprehensive data protection and it suits both structured and unstructured data.

C. Limitations of DLPD Techniques

Data Loss Prevention (DLP) is a powerful cybersecurity technique, but like any technology, it has its limitations and challenges. Here are some common limitations of DLP:

False Positives and Negatives: DLP solutions can generate false positives (incorrectly flagging legitimate activity as a breach) and false negatives (failing to detect actual breaches). Achieving the right balance between accuracy and alert volume can be challenging and may require ongoing tuning.

Encryption and Encrypted Traffic: DLP may struggle to inspect and monitor encrypted data traffic, as it cannot see the contents of encrypted files or messages. While some DLP solutions support integration with encryption technologies, they may not provide complete coverage.

Data Classification Complexity: Properly classifying and tagging sensitive data can be a complex and time-consuming task. Organizations need to define what constitutes sensitive data and create accurate policies to protect it.

Endpoint Limitations: DLP may have limitations when it comes to protecting data on endpoints (e.g., laptops, mobile devices). It can be challenging to enforce policies consistently on devices that are not connected to the corporate network.

D. Challenges of DLPD Techniques

Data Loss Prevention (DLP) techniques face several challenges, including the difficulty of striking the right balance between security and usability, the potential for generating false positives and negatives, the complexity of accurately classifying and tagging sensitive data, struggles with monitoring encrypted traffic and zero-day threats, the resource-intensive nature of implementation, and concerns related to privacy and employee acceptance. Additionally, DLP may not effectively address insider threats or malicious actors who have legitimate access to data, requiring organizations to adopt a multi-layered security approach to enhance overall data breach prevention and detection.

III. CONCLUSION

Preventing and detection of data leaks requires continuous efforts and investment from the organizations. In this paper, presented a review of data leak threat types and its techniques. Data Loss Prevention (DLP) techniques play a critical role in safeguarding sensitive information and preventing data breaches within organizations. These techniques encompass a diverse range of methods, each tailored to address specific security challenges. Whether it's through content scanning, contextual analysis, exact file matching, or statistical approaches, DLP techniques collectively work to protect valuable data assets. By implementing and continuously improving DLP techniques, organizations can mitigate risks, maintain trust with customers, and safeguard their confidential information in an ever-evolving digital landscape. Employees receive cyber security training only during on-boarding. Analysis is that Regular, ongoing training sessions need to be conducted to keep employees updated about emerging threats and best practices.

REFERENCES

- [1] Cheng, L., Liu, F., & Yao, D. (2017). Enterprise data breach: causes, challenges, prevention, and future directions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(5), e1211.
- [2] Fleury-Charles, A., Chowdhury, M. M., & Rifat, N. (2022, May). Data Breaches: Vulnerable Privacy. In *2022 IEEE International Conference on Electro Information Technology (eIT)* (pp. 538-543). IEEE.
- [3] AlKilani, H., Nasereddin, M., Hadi, A., & Tedmori, S. (2019, December). Data exfiltration techniques and data loss prevention system. In *2019 International Arab Conference on Information Technology (ACIT)* (pp. 124-127). IEEE.
- [4] Alneyadi, S., Sithirasanen, E., & Muthukkumarasamy, V. (2016). A survey on data leakage prevention systems. *Journal of Network and Computer Applications*, 62, 137-152.
- [5] Shu, X., Yao, D., & Bertino, E. (2015). Privacy-preserving detection of sensitive data exposure. *IEEE transactions on information forensics and security*, 10(5), 1092-1103.
- [6] Securosis, L. L. C. (2010). Understanding and selecting a data loss prevention solution. *Securosis, LLC*.
- [7] Takebayashi, T., Tsuda, H., Hasebe, T., & Masuoka, R. (2010). Data loss prevention technologies. *Fujitsu Scientific and Technical Journal*, 46(1), 47-55.
- [8] High-profile Company Data Breaches 2023, <https://www.electric.ai/blog/recent-big-company-data-breaches>
- [9] Tahboub, R., & Saleh, Y. (2014, January). Data leakage/loss prevention systems (DLP). In *2014 World Congress on Computer Applications and Information Systems (WCCAIS)* (pp. 1-6). IEEE.
- [10] 2016 cost of data breach study: global analysis. <https://www.cloudmask.com/hubfs/IBMstudy.pdf>
- [11] Shu, X., Yao, D., & Bertino, E. (2015). Privacy-preserving detection of sensitive data exposure. *IEEE transactions on information forensics and security*, 10(5), 1092-1103..
- [12] Shabtai, A., Elovici, Y., & Rokach, L. (2012). *A survey of data leakage detection and prevention solutions*. Springer Science & Business Media.
- [13] Yu, Y. (2022). Encryption Technology for Computer Network Data Security Protection. *Security and Communication Networks*, 2022.
- [14] Zhang, R., & Hu, Z. (2021). Access control method of network security authentication information based on fuzzy reasoning algorithm. *Measurement*, 185, 110103.
- [15] Borky, J. M., Bradley, T. H., Borky, J. M., & Bradley, T. H. (2019). Protecting information with cybersecurity. *Effective Model-Based Systems Engineering*, 345-404.
- [16] Hijji, M., & Alam, G. (2022). Cybersecurity Awareness and Training (CAT) Framework for Remote Working Employees. *Sensors*, 22(22), 8663.
- [17] Anderson, C. L., & Agarwal, R. (2010). Practicing safe computing: A multimethod empirical examination of home computer user security behavioral intentions. *MIS quarterly*, 613-643.