WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Survey on Data Leakage Detection and Prevention

Nikhil D Gowda[1], Murugan R[2]

[1] 2nd Year MCA(ISMS) School of Computer Science and IT Jain (Deemed-to-be-University), Bengaluru, India
[2]Programme Head-MCA, Dept. of CS & IT, Jain (Deemed-to-be-University), Bengaluru, India
DOI; https://doi.org/10.55248/gengpi.5.0324.0701

ABSTRACT:

Many firms today move their company from one level to another. Through a variety of companies and agencies, they do business. The storage and transfer of data from one location to another rose along with the company level, increasing the possibility of data leakage by any user in the midst. Finding the guilty agent who leaked the sensitive material is our first goal. The term "information leakage" describes the unapproved dissemination of data from a server area to the public. The appropriation model for data leakage counteractive action is described in those publications, and a document allocation plan with minimum cover between the client sets of records is selected. This allows for a high likelihood of detecting leaked sources. Sensitive information in firms, such as internal regulations, financial data, and personal information, might be compromised by a hostile user. Information leaking is a big concern for a variety of businesses. Some data leakage detection models employed 'fake objects' stored in the server database. The fabricated items aid in identifying the user who leaked the material. The guilt probability refers to the likelihood of each user leaking the file.

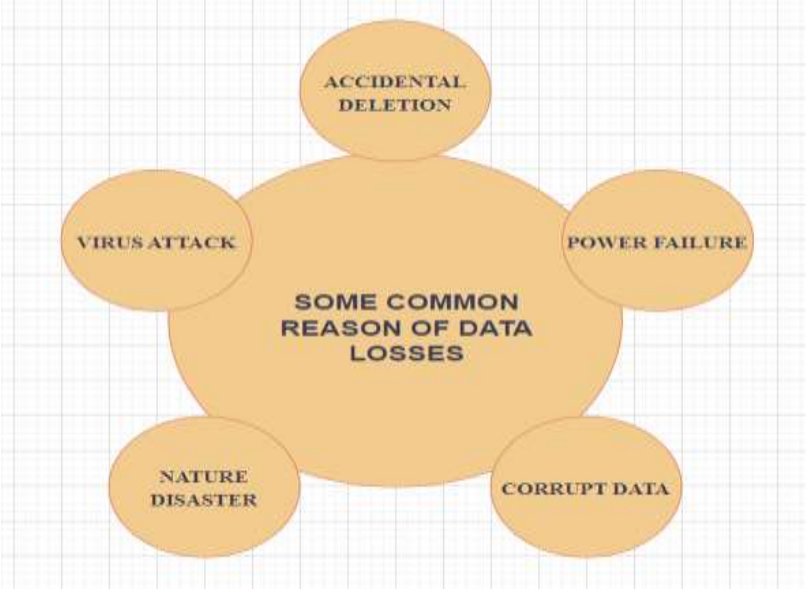
KEYWORDS: *Fake object, Guilt probability, Minimum Overlap, Distribution Model, and Cloud Computing*

I. INTRODUCTION

Cloud computing has led to increased data storage, but ensuring data security has become increasingly important. If data is disclosed, the firm may suffer significant losses. To detect data leaking, some models added bogus items to the collection of users. The agents see a phony object as genuine, despite its absence from reality. The data allocation approach includes a false object in the distribution set with sensitive data. The fake object serves as a watermark to identify the correct owner. If sensitive data is released, a phony item might assist identify the source of the leak. Sensitive data can be leaked and discovered in unauthorized areas. An organization may need to exchange consumer information with a partner organization. Data distribution to an illegitimate firm might lead to leaks. Sensitive data might include information about clients, managers, patients, and other authorized users. According to public data leakage reports, there are many sorts of data leakages and their percentages.

II. REASON FOR DATA LOSSES

There are some of common reason of data losses



2.1 Accidental deletion:

Data can sometimes be mistakenly removed from your storage disk. That data is not noticed by the user for an extended period.

2.2 Power failure:

When working on electrical equipment like laptops or PCs, turning them off unexpectedly increases the risk of losing sensitive data. To avoid this problem, save your work often.

2.3 Corrupt Data :

If the database is damaged, there is a risk of data loss. With the correct tools, data may be recovered from a corrupted file.

2.4 Nature Disaster:

Assume your file database is affected by natural disasters like fires, floods, and other unanticipated events. We can overcome this issue by storing the data in another location.

2.5 Virus Attack :

When a system is infected with a hidden virus, it attacks the database, causing data corruption.

III. PRIVILEGES

3.1 Overlapping the data:

The distributor assigns the data to an agent with a constraint and aim. The constraint is to complete the agent's request for the file. The distributor manages the table for user requests. The objective can identify the leak. The distributor selects the shortest way to distribute the file to the user who requested it. The model relied on a single file transaction to satisfy the client. If a file is sent from A to B via just two mediators (X and Y), the likelihood of the file being leaked during the transaction is (½). File leakers can be either X or Y. If we do not apply the least overlap approach and transmit the identical file from A source to B destination in between, there can be N mediators (n1, n2, …, nN). If the file is leaked during the transaction process, the likelihood becomes (1/n), making it more difficult to identify rogue users. So, least overlap is an effective way to locate the guilty user.

3.2 Fake object :

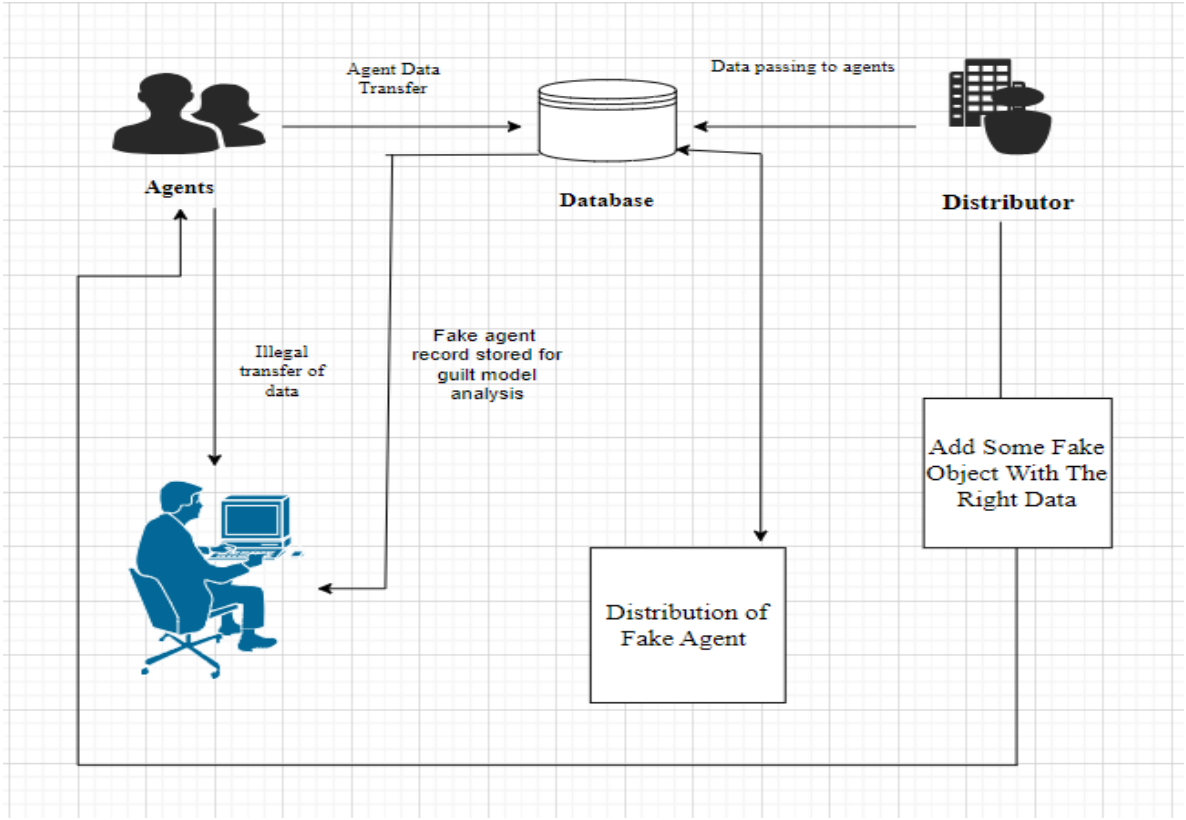
The distributor creates a fake item. To boost the efficiency of detecting the leaking source, the distributer included some bogus objects alongside the genuine data. The thing isn't real. It is only created during the transaction to identify the guilty agent who is meant to leak the data. The false object is constructed in such a way that the agent is unable to tell the difference between it and a real thing. Fake records are typically used to track or monitor actual data. For example, if X company sells an item to company Y, company X adds some fake tracing software that includes their company address. If

company Y misuses their data to sell outsiders their data, company X automatically receives a copy of the issue records, allowing company X to easily identify the improper use of data.

3.3 Guilt Assessment :

The data or file leak or loss caused by unauthorized clients or users, symbolized by the letter L. Clients who have part of the leaked data of L may be highly sensitive to releasing information or data. Following the file leak, he may claim that he is innocent and that the L data were obtained from the distributor and sent to the target users or organization in another method. Our goal is to identify the people or clients who leak the files the most. For example, if any of L's objects or data were transferred to just agent A1, we could suspect A1 more since only A1 users had access to the file set. So, the likelihood of that customer A1 is Guilt leak likelihood.

IV. PROCESS OF DATA LEAKAGEDETECTION



1. The distributor logins into the system.
2. The distributor uploads the Data [example. text files] into the Database.
3. Agent asks for the file or distributor uploads all file for agents accordingly along with private key after Login into the system.
4. The distributor sends that requested file to the requested agents who add some fake objects.
5. Agents will download the files according to his needs [Sample requests or explicit request].
6. If any agents leak the data to the third party [Fake Agents] the distributor will check for the leaked data and will find the file which has been leaked

4.1 Encryption Techniques

Encryption is the act of transforming plaintext data into ciphertext using mathematical methods and cryptographic keys, making it unreadable without the accompanying decryption key.

. Symmetric Encryption: Symmetric encryption employs a single key for encryption and decoding.

AES, DES, and Triple DES are all common encryption techniques. Symmetric encryption is quick and effective, but it requires safe key management to prevent unwanted access.

. Asymmetric Encryption: Asymmetric encryption uses public and private keys for encryption and decryption, respectively.

Examples include RSA (Rivest-Shamir-Adleman) and Elliptic Curve Cryptography (ECC).

Asymmetric encryption increases security and allows for safe communication between parties, but it is computationally costly.

4.2 Tokenization Techniques:

Tokenization replaces sensitive data with non-sensitive tokens that are produced at random and contain no significant information. Tokenization techniques are extensively used in payment processing, cloud computing and data storage.

. Format-Preserving Tokenization: Tokenization that preserves format and length preserves the original data while replacing it with a token.

It ensures interoperability with current systems and databases while minimizing the need for large data structure modifications.

. Random Tokenization: Random tokenization creates tokens from alphanumeric letters or cryptographic hashes.

It improves security by removing any relationship between tokens and original data.

. Secure Vaulting: Secure vaulting is the process of keeping sensitive data in a secure repository or vault while providing access tokens. It centralizes data protection while lowering the danger of data exposure.

V. PREVENTION OF DATALEAKAGE

5.1 Access Controls:

Access controls are a set of mechanisms that limit access to resources based on preset policies and permissions. These controls can be applied at a few levels, including physical, logical, and administrative. Common access control techniques include the following:

. Role-depending Access Control (RBAC): Users are granted permissions depending on their positions within an organization. This concept streamlines access control by assigning rights to predefined roles rather than individual individuals.

. Mandatory Access Control (MAC): MAC implements access regulations specified by system administrators or security policies. It limits users' ability to change access controls, giving a high level of security but necessitating careful setup and monitoring.

. Discretionary Access Control (DAC): DAC enables users to restrict access to resources that they own. Users can give or remove access to other users or groups, which provides flexibility but may expose security issues if not managed effectively.

5.2 User Authentication Mechanisms:

User authentication validates the identification of people seeking to access a system or resource. Authentication systems employ a variety of elements to confirm user identities, including as

. Password-based Authentication: Passwords are the most used type of authentication, requiring users to enter a unique string of characters in order to access their accounts. However, passwords are vulnerable to brute force assaults, phishing, and password guessing.

. Multi-Factor Authentication (MFA) improves security by requiring users to give several authentication methods, such as passwords, biometric data, or one-time codes delivered to their mobile devices. This strategy considerably decreases the likelihood of illegal access.

. Biometric authentication is the use of unique biological traits, such as fingerprints, facial features, or iris patterns, to authenticate the identity of users. Biometric data is difficult to reproduce and provides a high level of security and convenience.

5.3 Data loss prevention (DLP) solutions:

Data Loss Prevention (DLP) solutions are a collection of technologies and tactics that work together to prevent the unauthorized exposure or leaking of sensitive information. Data Discovery and Classification DLP systems use powerful scanning and classification algorithms to detect sensitive data stored in a variety of repositories, including databases, file servers, and cloud storage platforms. Organizations can prioritize protection efforts and implement suitable security rules by classifying data depending on its sensitivity level. Policy-Based Enforcement DLP systems allow enterprises to set detailed security policies that control the handling and transfer of sensitive data. These rules may include data sharing limits, encryption standards, and monitoring of data access and use. Policy enforcement tools guarantee that data protection procedures are implemented uniformly across the company. Content Inspection and Contextual Analysis: DLP systems use content inspection techniques such as keyword matching, regular expressions, and machine learning algorithms to assess data transactions' context and purpose. DLP systems can properly detect possible security threats and take necessary corrective steps by analyzing aspects such as user behavior, data content, and contextual information. issue Response and Remediation when a policy violation or security issue occurs, DLP systems provide real-time warnings and notifications to security administrators. These signals may prompt automatic reaction actions

such as data transfer stopping, file quarantining, or forensic inquiry. Prompt incident response and cleanup techniques are crucial for mitigating the effect of data breaches and maintaining regulatory compliance.

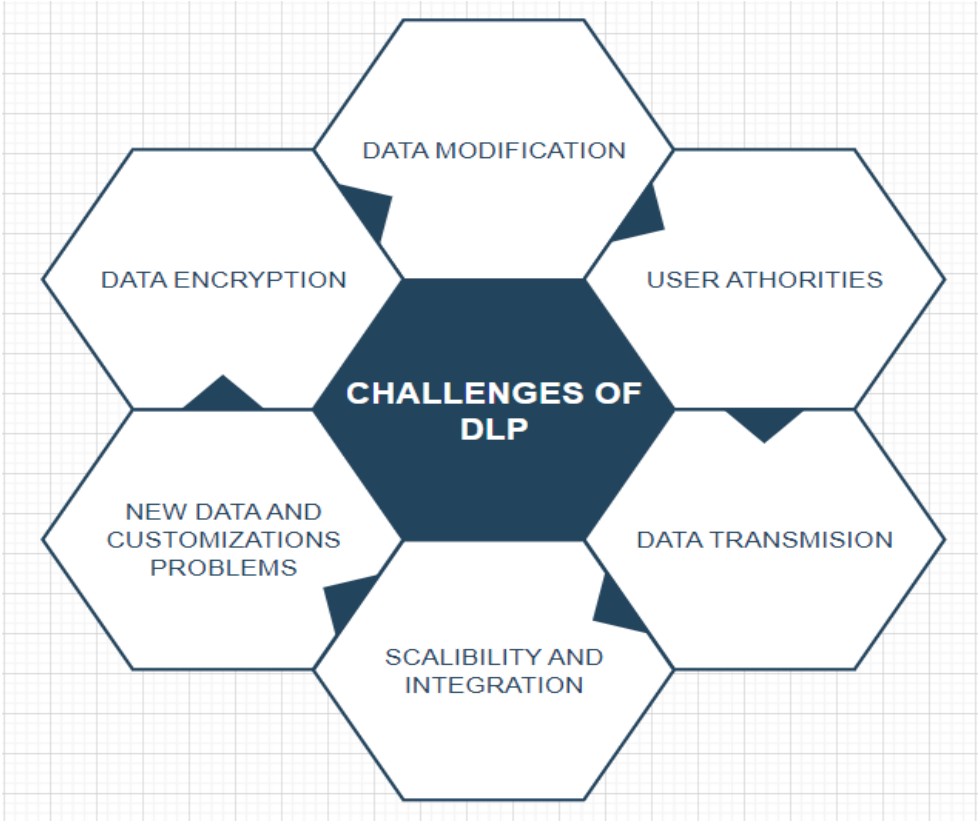
5.4 Significance of DLP Solutions:

Data Loss Prevention (DLP) solutions provide numerous advantages to organizations seeking to safeguard their sensitive information assets Prevent Data Breach: DLP systems assist companies in preventing data breaches by proactively identifying and mitigating security risks such as unauthorized data access, exfiltration, or insider threats. Ensure Regulatory Compliance DLP systems help firms comply with industry rules, data protection legislation, and privacy regulations by enforcing security policies, monitoring data flows, and keeping audit trails. Protect Intellectual Property: By protecting sensitive intellectual property and trade secrets, DLP systems help enterprises to maintain their competitive edge and innovation.

Maintain Reputational Integrity Successful DLP implementation builds consumer trust and confidence by demonstrating a commitment to data protection, privacy, and compliance. Protecting sensitive information from unauthorized exposure or misuse helps businesses maintain their reputation and brand value.

VI. CHALLENGES

With the increased popularity of data leakage prevention/detection, numerous significant issues have arisen. These issues often develop during data storage and transmission from one node to another.



6.1 Data encryption

The encryption approach can safeguard data from a hostile user. The encryption technology allows us to achieve integrity, secrecy, and authenticity. However, if a powerful encryption technology is applied, the data file will be tough to examine.

6.2 Data modification

When data is updated, it might automatically be partially disclosed to users. The administrator faces a significant issue in securing their data during alteration.

6.3 User authorities

In current times, there is a lot of data being stored in the cloud, but only the authorized users may access it. If there are no restrictions on users, any user can access all of the data, therefore there are certain issues when keeping data in the cloud. To strengthen security, the administrator grants unique privileges to each user to view their file.

6.4 Data transmission

Data may be conveyed in a variety of ways; if the data is transmitted via a dedicated or specialized channel, it maintains confidentiality; however, if the channels are other than specified ones, such as USB, email, or another format, it becomes difficult to safeguard the data.

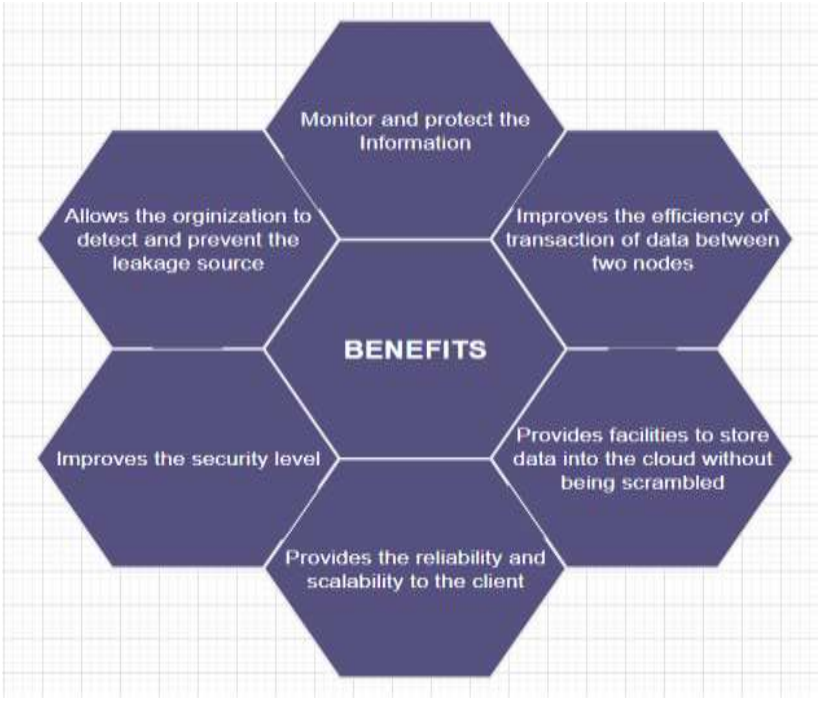
6.5 New Data and customization

Every user may have a varied set of data formats. Sometimes it is unsuitable for cloud storage, making it tough for administrators to maintain their data. Whenever new user information is saved in the cloud, the user must submit all document-related information.

6.6 Scalability and integration

These days, the large firm has a limitless quantity of data which they normally want to save their data into the cloud. This is a safe approach to protect their data but also have their downsides in which data size is enormous so that monitoring, matching, and obtaining their data becomes more challenging.

VII. BENEFITS



7.1 DLP innovation enables ensure to user and control of all types of information, including client information, monetary information, and protected innovation. With the help of data leakage prevention, we protect the data from unauthorized users.

7.2 Data Leakage Prevention detects classified information, ensures that it does not enter the cloud without being scrambled, and is only provided to cloud apps that have been approved.

7.3 DLP enables the business to identify and limit data leakage by stopping it, so avoiding future danger.

7.4 Data leakage approaches can strengthen security principles in a variety of ways, including look-based data, guilt likelihood, and fake objects.

7.5 Data Breach Prevention DLP solutions assist to prevent unauthorized access, leakage, or theft of sensitive information, lowering the risk of data breaches and the financial and reputational harm that comes with them. Secure Intellectual Property DLP solutions preserve enterprises' competitive edge and innovative capabilities by preserving sensitive intellectual property, trade secrets, and private information. Ensure Data Integrity: DLP systems contribute to the integrity of important data assets by enforcing security regulations, preventing unauthorized changes, and identifying tampering or data manipulation efforts.

7.5 Regulatory Compliance Address Data Privacy requirements: DLP solutions help firms comply with data protection requirements like GDPR, HIPAA, PCI DSS, and CCPA by enforcing security rules, monitoring data transfers, and preserving audit trails. Implementing DLP solutions demonstrates an organization's commitment to data security and regulatory compliance, lowering the probability of fines, penalties, and legal liabilities associated with noncompliance.

Here, the following enumerates some of the most renowned web-site builders.

7.6 Risk Mitigation Minimize Insider risks DLP systems assist to minimize insider risks by monitoring user activity, identifying aberrant actions, and enforcing access rules to prevent illegal data access, abuse, or exfiltration by employees, contractors, or trusted partners. External Threat Mitigation DLP systems guard against external threats such as malware, ransomware, phishing assaults, and advanced persistent threats (APTs) by identifying and stopping harmful activity aimed at stealing sensitive data or compromising corporate assets.

7.6 Operational Efficiency DLP systems improve data governance by giving insight into data flows, usage patterns, and access rights, allowing companies to discover and address security threats, compliance gaps, and operational inefficiencies. Streamline Incident Response: DLP systems simplify incident response procedures by offering real-time warnings, automated remedial actions, and forensic analysis capabilities, allowing businesses to respond quickly to security problems and minimize their impact on business operations.

7.7 Protect Brand Reputation DLP solutions increase consumer trust and confidence in the organization's capacity to preserve their personal and private data, sustaining brand reputation and loyalty. Avoid Reputational Damage Data loss prevention solutions assist firms in avoiding reputational damage caused by data breaches, privacy violations, or security events, which can result in unfavorable publicity, loss of consumer trust, and long-term brand reputation harm.

7.8 Cost Savings Reduce Financial Losses DLP solutions assist firms in avoiding financial losses caused by data breaches, regulatory fines, legal expenses, and remediation costs, resulting in considerable long-term cost savings.

DLP systems increase operational efficiency and cost-effectiveness by automating data protection operations, improving resource allocation, and reducing the impact of security events.

VIII. CONCLUSION

In certain data leakage detection models, we assume a restricted number of users, and all actions are done between these users. This issue is tied to the cloud, thus in the future, anybody can submit a request to the cloud for a specific file that is not already in the user database. We are implementing some new features that consider dynamic users, which means that any number of people can request the content. The expansion of the data allocation approach so that agent queries can be handled online. The Distributor inserts some extra false objects during the allocation of data to increase the security of the data. These days the data leakage becomes a serious and silent issue for the organization. Data leakage occurs from different sources which mainly people. This uncontrollable information leakage placed commercial enterprise in a susceptible position. Once this information is now not in the area, Then the company is at severe risk. This sensitive data can be electronically distributed through email, web pages, spreadsheets, USB keys, and other electronic devices.

Nowadays, the company is not restricted to a specific place but rather spreads around the world. So, suppose one customer delivers data from one nation to another via certain agents. During the transaction, anyone might leak data. There are methods and strategies for detecting guilty users and preventing leaks by blocking their sources. In this study, work on basically two aspects: the first is false objects included in-database, and the second is data allocation technique to distribute the data to customers with the smallest transaction.

The survey results highlight the necessity of a multifaceted strategy to data leakage prevention that combines effective detection techniques with proactive actions. Organizations may improve their capacity to identify, mitigate, and prevent data leakage across a range of contexts and attack vectors by employing modern technologies such as artificial intelligence, encryption. Furthermore, this study highlighted current issues and upcoming trends in data leakage detection and prevention. As data breaches become more sophisticated and widespread, there is a greater demand for creative solutions that can react to changing threat landscapes and emerging attack vectors. Furthermore, the growing use of cloud services, mobile devices, and remote work arrangements demands the creation of comprehensive data leakage prevention techniques that meet the issues presented by these settings.

VIII. REFERENCES

[1] Yin Fan, Wang Lina, Yu Rongwei, Ma Xiaoyan"A Distribution model for Data Leakage Prevention," 2013 IEEE International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC 2013), Shenyang, China.

[2] S. Praveen Kumar, Y. Srinivas, D. Suba Rao, Ashish Kumar, "A Novel Model for Data Leakage Detection and Prevention in Distributed Environment," 2016 International Journal of Engineering and Technical Research (IJETR).

[3] K. Kaur, I. Gupta and A. K. Singh, "Data Leakage Prevention: Email Protection via Gateway", J. Phys.: Conf. Ser., vol. 933, no. 1, IOP Publishing, 2018.

[4] .Papadimitriou P, Garcia-Molina H, "A Model for Data Leakage Detection," 2011 IEEE Transaction on Knowledge and Data Engineering.

[5] I. Gupta and A. K. Singh, "A Probabilistic Approach for Guilty Agent Detection using Bigraph after Distribution of Sample Data", Procedia Computer Science, vol. 125, pp. 662-668, 2018.

[6] U. Arora, S. Verma, I. Gupta and A. K. Singh, "Implementing privacy using modified tree and map technique", 3rd International Conference on Advances in Computing, Communication & Automation (ICACCA), pp. 1-5, IEEE, 2017.

[7] I. Gupta and A. K. Singh, "Dynamic Threshold based Information Leaker Identification Scheme", Information Processing Letters, vol. 147, pp. 69-73, 2019.

[8] Animesh Nag, Anand Kesharwani, Abhishek Tiwari, Ishu Gupta, Bharti Sharma, Ashutosh Kumar Singh, “Potential and Extension of Comparative Internet of Things”, 2nd International Conference on Computer Networks and Inventive Communication Technologies, Coimbatore, India, 2019.

[9] R. K. R. Chandran and P. Kanagasabai, "A comprehensive survey on data leakage detection and prevention techniques," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018.

[10] S. Vishalini and N. Dharani, "A survey on data leakage detection and prevention," in 2015 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2015.