International Journal of
Computational Learning & Intelligence
Vol. 03, Issue 02 (April) 2024

RESEARCH ARTICLE OPEN ACCESS

Data Leakage and Detection

Prathik B K . Manovikyath L . Sandeep Y . Aryaman . Abhiram P

School of Computer Science and Engineering
REVA University, Bengaluru, India

DOI: **10.5281/zenodo.11065679**
Received: 21 January 2024 / Revised: 11 February 2024 / Accepted: 12 April 2024

Abstract – We study the following problem: A data distributor has given sensitive data to a set of supposedly trusted agents (third parties). Some of the data are leaked and found in an unauthorized place (e.g., on the web or somebody's laptop). The distributor must assess the likelihood that the leaked data came from one or more agents, as opposed to having been independently gathered by other means. We propose data allocation strategies (across the agents) that improve the probability of identifying leakages. These methods do not rely on alterations of the released data (e.g., watermarks). In some cases, we can also inject "realistic but fake" data records to further improve our chances of detecting leakage and identifying the guilty party "realistic but fake" data records to further improve our chances of detecting leakage and identifying the guilty party.

Index Terms – Campus Network, Cisco Packet Tracer, Network Design, VLAN, DHCP, Quality of Service, Network Security, Wireless Integration, Management and Monitoring.

I. INTRODUCTION

Data leakage is defined as the accidental or unintentional distribution of private or sensitive data to an unauthorized entity .Data leakage poses a serious issue for companies as the number of incidents and the cost to those experiencing them continue to increase. Data leakage is enhanced by the fact that transmitted data including emails, instant messaging, website forms, and file transfers among others, are largely unregulated and unmonitored on their way to their destinations. The main scope of this module is providing complete information about the data/content that is accessed by the users within the website. Forms Authentication technique is used to provide security to the website in order to prevent the leakage of the data. Continuous observation is made automatically and the information is send to the administrator so that he can identify whenever the data is leaked.

In the course of doing business, sometimes sensitive data must be handed over to supposedly trusted third parties. For example, a hospital may give patient records to researchers who will devise new treatments. Similarly, a company may have partnerships with other companies that require sharing customer data. Another enterprise may outsource its data processing, so data must be given to various

other companies. We call the owner of the data the distributor and the supposedly trusted third parties the agents. Our goal is to detect when the distributor's sensitive data has been leaked by agents, and if possible to identify the agent that leaked the data. We consider applications where the original sensitive data cannot be perturbed. Perturbation is a very useful technique where the data is modified and made "less sensitive" before being handed to agents. For example, one can add random noise to certain attributes, or one can replace exact values by ranges However, in some cases it is important not to alter the original distributor's data. For example, if an outsourcer is doing our payroll, he must have the exact salary and customer bank account numbers. If medical researchers will be treating patients (as opposed to simply computing statistics), they may need accurate data for the patients. Traditionally, leakage detection is handled by watermarking, e.g., a unique code is embedded in each distributed copy. If that copy is later discovered in the hands of an unauthorized party, the leaker can be identified. Watermarks can be very useful in some cases, but again, involve some modification of the original data. Furthermore, watermarks can sometimes be destroyed if the data recipient is malicious.

II. LITERATURE SURVEY

Data leakage poses significant threats to organizations, leading to breaches in confidentiality, integrity, and availability of sensitive information. Detecting and preventing data leakage have become paramount in safeguarding valuable assets. This literature survey aims to explore the diverse landscape of data leakage detection methodologies, highlighting key approaches, challenges, and emerging trends. Methods of Data Leakage Detection: Various approaches have been proposed for data leakage detection, ranging from rule-based methods to machine learning and AI-driven techniques. Rule-based systems rely on predefined policies and signatures to identify potential data leaks. However, they often struggle with detecting unknown or evolving threats. On the other hand, machine learning models, such as anomaly detection and pattern recognition, have shown promise in detecting previously unseen data leakage patterns by learning from historical data. Technologies and Tools: Several technologies and tools have been developed to aid in data leakage detection. Encryption and tokenization techniques help protect data at rest and in transit. Data Loss Prevention (DLP) solutions, employing content inspection and contextual analysis, aim to prevent unauthorized data exfiltration. Additionally, advancements in data masking and obfuscation technologies contribute to mitigating the risks associated with data leakage.

Challenges and Limitations: Despite advancements, challenges persist in effectively detecting data leakage. One significant challenge lies in handling the vast volume and variety of data generated across diverse platforms and devices. Additionally, distinguishing between legitimate data access and potential leaks without hindering operational efficiency remains a challenge. Moreover, emerging technologies like edge computing and IoT devices introduce new complexities in data leakage detection strategies. Emerging Trends and Future Directions: The future of data leakage detection appears promising with the integration of advanced technologies. The convergence of AI, particularly deep learning, and Big Data analytics holds potential for more accurate and proactive detection methods. Furthermore, the adoption of Zero Trust architectures, where no entity is inherently trusted, coupled with continuous monitoring, is gaining traction in preventing data leakage.

III. PROPOSED WORK

- Data Classification: Classify data based on sensitivity levels to identify critical information.
- Access Controls: Implement robust access controls to limit data access to authorized personnel.
- Monitoring Tools: Employ monitoring tools to track data access and detect unusual patterns.
- Encryption:Encrypt sensitive data during storage and transmission to protect it from unauthorized access.
- User Behavior Analytics:Utilize user behavior analytics to detect abnormal activities that may indicate data leakage.
- Endpoint Security:Strengthen endpoint security to prevent unauthorized access from devices connected to the network.
- Regular Audits:Conduct regular audits to assess the security posture and identify vulnerabilities.
- Incident Response Plan: Develop a comprehensive incident response plan to address data leakage incidents promptly.
- Employee Training: Educate employees on security best practices and the importance of safeguarding sensitive information.
- Data Loss Prevention (DLP) Software: Implement DLP solutions to monitor, detect, and prevent unauthorized data transfers.
- Network Segmentation:Segment the network to contain potential breaches and limit lateral movement within the system.
- Anomaly Detection: Utilize anomaly detection algorithms to identify deviations from normal data access patterns.
- Collaboration with Cybersecurity Experts: Collaborate with cybersecurity experts to stay updated on the latest threats and mitigation strategies.
- Legal and Regulatory Compliance:Ensure compliance with data protection laws and industry regulations to avoid legal consequences.
- Regular Updates and Patching:Keep all software and systems up-to-date with the latest security patches to address vulnerabilities.

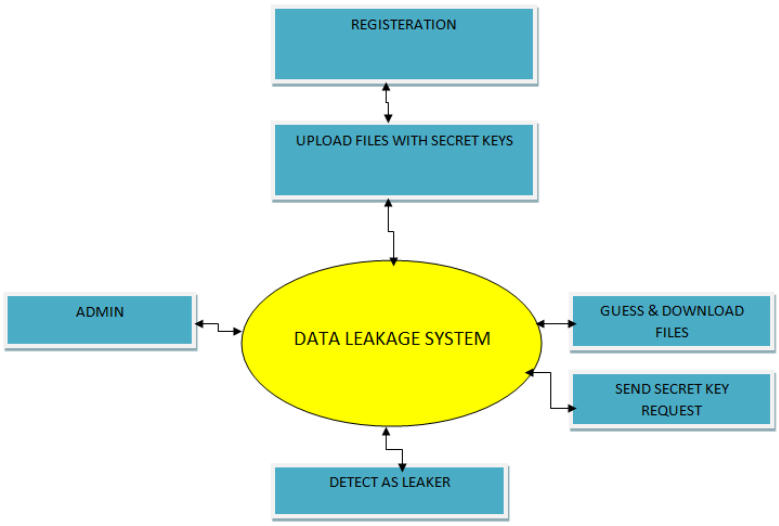

Fig. 1. Flow of proposed work

IV. IMPLEMENTATION

Sr No	Request By	File	Confirm	Date_time
1	user1	**Subject:** excel notes **File Name:** 6048ff4e8cb07aa60b6777b6f7384d52-LEAVESYSTEM.xlsx **File Size:** 0.01Mb	Confirmed	2020-06-05 09:29:57

Fig. 2. Showing key request table

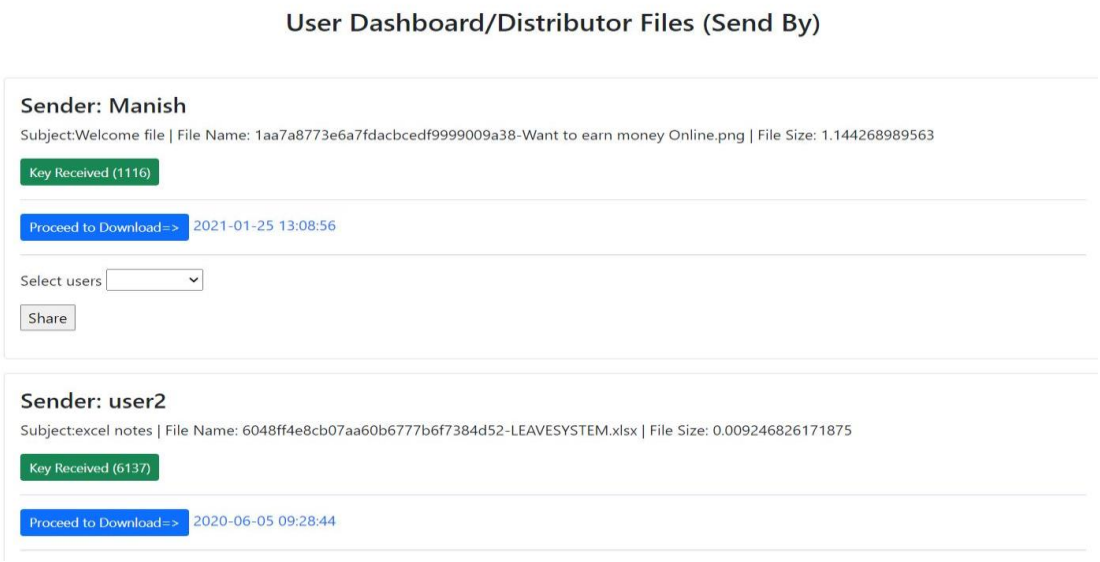

Fig. 3. files send by other users

Sr No	User Name	Email Id	Activate account	Create Distributor
1	Manish	manish@gmail.com	Deactivate	Make Distributor
2	user2	user2@gmail.com	Deactivate	Make Distributor
3	user1	user1@gmail.com	Deactivate	Make Distributor
4	distributor	distributor@gmail.com	Deactivate	Make User

Fig. 4. User registration list for admin

Sr No	File Details	Download	Shared By	Date Time
1	**Subject:** Welcome file **File Name:** 1aa7a8773e6a7fdacbcedf9999009a38-Want to earn money Online.png **File Size:** 1.14Mb	Download (1116)	admin	2021-01-25 13:11:41
2	**Subject:** excel notes **File Name:** 6048ff4e8cb07aa60b6777b6f7384d52-LEAVESYSTEM.xlsx **File Size:** 0.01Mb	Download (6137)	distributor	2020-06-05 09:33:33

Fig. 5. Showing leakers list

V. CONCLUSION

Data leakage is becoming a nuisance in many organizations because of the increasing incidents that can result in the challenge. Thus, companies need to make sure that all the stakeholders understand about the challenge since it is a business wide challenge. The reason for that is to help the people understand that the corporate security is critical, and they also need to understand the procedures as well as the policies that are helpful in achieving a secure environment. There should be the usage of technologies that can detect and prevent the data leakage problems.

REFERENCES

1. Papadimitriou, P., & Garcia-Molina, H. (2011). Data leakage detection. Knowledge and Data Engineering, IEEE Transactions on,23(1), 51-63.
2. Shabtai, A., Elovici, Y., & Rokach, L. (2012).A survey of data leakage detection and prevention solutions.
3. Ahmed, S. T., Kumar, V. V., & Kim, J. (2023). AITel: eHealth augmented-intelligence-based telemedicine resource recommendation framework for IoT devices in smart cities. *IEEE Internet of Things Journal*, *10*(21), 18461-18468.
4. Sathiyamoorthi, V., Ilavarasi, A. K., Murugeswari, K., Ahmed, S. T., Devi, B. A., & Kalipindi, M. (2021). A deep convolutional neural network based computer aided diagnosis system for the prediction of Alzheimer's disease in MRI images. *Measurement*, *171*, 108838.
5. Ahmed, S. T., Basha, S. M., Ramachandran, M., Daneshmand, M., & Gandomi, A. H. (2023). An edge-AI enabled autonomous connected ambulance route resource recommendation protocol (ACA-R3) for eHealth in smart cities. *IEEE Internet of Things Journal*.
6. Springer Science & Business Media. Wu, J., Zhou, J., Ma, J., Mei, S., & Ren, J. (2011).
7. In Intelligence Information Processing and Trusted Computing (IPTC), 2011 2nd International Symposium on(pp. 39-42). IEEE.