SIMATS SCHOOL OF ENGINEERING

SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES

CHENNAI-602105

DATA LEAKS DETECTION SYSTEM

A CAPSTONE PROJECT REPORT

Submitted in the partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE ENGINEERING

Submitted by

Harshana B(192211067)

Under the Guidance of

Dr. Antony Joseph Rajan D

June 2024

DECLARATION

I am Harshana B, student of **Bachelor of Engineering in Computer Science Engineering,** Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, hereby declare that the work presented in this Capstone Project Work entitled **Data Leaks Detection System** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics.

(Harshana B(192211067))

Date:

Place:

CERTIFICATE

This is to certify that the project entitled **“Data Leaks Detection System”** submitted by **Harshana.B(192211067)** has been carried out under our supervision. The project has been submitted as per the requirements for the award of degree.

Project Supervisor

Table of Contents

S.NO	TOPICS
	Abstract
1.	**Introduction**
2.	**Existing System**
3.	**Literature survey**
4.	**Proposed System**
5.	**Implementation**
6.	**Conclusion & Future Scope**

ABSTRACT:

Data leaks pose a significant threat to organizational security, leading to the unauthorized disclosure of sensitive information. A Data Leaks Detection System (DLDS) is designed to identify and mitigate these leaks by continuously monitoring data flows and access patterns. This system employs a combination of advanced techniques, including anomaly detection, machine learning algorithms, and real-time data analytics, to detect unusual activities that may indicate a potential leak. By integrating with various data sources such as network traffic, file systems, databases, and cloud services, the DLDS provides comprehensive coverage and timely alerts. Additionally, it incorporates encryption, access controls, and regular audits to enhance data security. The system's effectiveness is further bolstered by its ability to adapt to evolving threats through ongoing updates and the incorporation of new threat intelligence. Overall, a robust DLDS not only helps in early detection and response to data breaches but also strengthens the overall data governance framework, ensuring compliance with regulatory requirements and protecting organizational assets.

In the evolving landscape of cloud computing, efficient resource utilization and allocation are paramount to ensuring optimal performance and cost-efficiency. This paper presents a novel approach to determine and prevent data leaks. By leveraging machine learning algorithms and historical data, our method forecasts the demand for various cloud resources, enabling dynamic and predictive scaling of cloud nodes.

1.INTRODUCTION:

Data breaches, especially those resulting from SQL injection attacks, pose a severe threat to the confidentiality, integrity, and availability of sensitive information. SQL injection is a type of cyber attack that manipulates a standard SQL query to exploit non-validated input vulnerabilities in a database-driven web application. This can lead to unauthorized access to confidential data, such as personal identification details, credit card information, and other sensitive records.

In this context, the need for advanced predictive and matchmaking techniques has become evident. The increasing reliance on digital transactions necessitates robust security measures to safeguard user data. Traditional security mechanisms often fall short in dynamically identifying and mitigating SQL injection attacks. Moreover, the storage of unencrypted data further exacerbates the risk, as attackers can access plain text information if they breach the security perimeters.

2.EXISTING SYSTEM:

Existing systems for data leaks detection encompass a range of tools and technologies designed to safeguard sensitive information from unauthorized access and exfiltration. These systems generally fall into several categories:

2.1. Data Loss Prevention (DLP) Solutions: DLP, or Data Loss Prevention, is a cybersecurity solution that detects and prevents data breaches. Since it blocks extraction of sensitive data, organizations use it for internal security and regulatory compliance.

- **Endpoint DLP**: Monitors and controls data transfers on endpoint devices (e.g., laptops, desktops) to prevent data leaks through external storage devices, email, or applications.

- **Network DLP**: Inspects data in transit over the network to identify and block sensitive information from being sent to unauthorized recipients.

- **Cloud DLP**: Extends data protection policies to cloud storage and SaaS applications, ensuring that sensitive data is not exposed through cloud services.

2.2. Intrusion Detection Systems (IDS) and Intrusion Prevention Systems (IPS)

An intrusion detection system is a device or software application that monitors a network or systems for malicious activity or policy violations. Any intrusion activity or violation is typically either reported to an administrator or collected centrally using a security information and event management system.

- **Signature-based IDS/IPS**: Utilizes predefined signatures of known threats to detect and prevent data leaks.
- **Anomaly-based IDS/IPS**: Identifies deviations from normal behavior that may indicate data leaks or malicious activity.

2.3. McAfee Total Protection for Data Loss Prevention

Some advanced systems are beginning to incorporate McAfee Total Protection for Data Loss Prevention. These systems use historical data and machine learning algorithms to forecast future resource demands and adjust allocations proactively.

- **Machine Learning Models:** By analyzing past usage patterns, these models can predict future resource needs with varying degrees of accuracy.

- **Hybrid Approaches:** Combining predictive models with reactive scaling can offer a more balanced solution, leveraging the strengths of both approaches. However, the integration and management of such hybrid systems can be complex and resource-intensive.

3.LITERATURE SURVEY:

Conducting a literature survey for " Data Leaks Detection System" involves reviewing existing research and methodologies in the fields of cloud computing, resource prediction, and matchmaking. Here's an organized overview of key topics and relevant literature

3.1. Data Loss Prevention (DLP) Solutions

- Highlights the importance of data classification, monitoring, and policy enforcement in preventing data leaks.
- **Key References:**
    - Takebayashi, T., Tsuda, H., Hasebe, T. and Masuoka, R., 2010. Data loss prevention technologies. Fujitsu Scientific and Technical Journal, 46(1), pp.47-55.
    - Costante, E., Fauri, D., Etalle, S., Den Hartog, J. and Zannone, N., 2016, May. A hybrid framework for data loss prevention and detection. In 2016 IEEE security and privacy workshops (SPW) (pp. 324-333). IEEE.

3.2. Intrusion Detection Systems (IDS)

- A widely-used open-source IDS/IPS, and its capabilities in detecting network-based attacks, which form a crucial component in identifying potential data leaks.
- **Key References:**
    - Sabahi, F. and Movaghar, A., 2008, October. Intrusion detection: A survey. In 2008 Third International Conference on Systems and Networks Communications (pp. 23-26). IEEE.
    - Khraisat, A., Gondal, I., Vamplew, P. and Kamruzzaman, J., 2019. Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecurity, 2(1), pp.1-22.

3.3. Machine Learning and Anomaly Detection

- **Key References:**
    - Chandola, V., Banerjee, A., & Kumar, V. (2009). "Anomaly Detection: A Survey." ACM Computing Surveys, 41(3), 1-58.
- **Summary**: This comprehensive survey explores various anomaly detection techniques applicable to data leaks detection in transactions on Parallel and Distributed Systems.

3.4. Security Information and Event Management (SIEM):

- **Key References:**
    - Scarfone, K., & Mell, P. (2007). "Guide to Intrusion Detection and Prevention Systems (IDPS)." NIST Special Publication 800-94.

- **Summary**: This guide outlines the role of SIEM systems in aggregating and analyzing log data from multiple sources to detect and respond to security incidents, including data leaks.

4.PROPOSED SYSTEM:

The proposed Data Leaks Detection System (DLDS) aims to enhance the detection and prevention of data leaks in organizational environments by utilizing advanced machine learning techniques, real-time monitoring, and comprehensive data analysis. The system is composed of several interconnected modules designed to work seamlessly to identify and mitigate potential data leaks.

4.1.Data Collection Module:

- **Objective**: Collect comprehensive data from various sources to monitor for potential data leaks.
- **Functionality**:
  - ❖ Gathers data from network traffic, endpoint devices, file systems, databases, and cloud services.
  - ❖ Collects information on user activities, access logs, data transfers, and system events.

4.2.Data Preprocessing Module:

- **Objective** Prepare the collected data for analysis and machine learning model training.

- **Functionality**:
  - ❖ Cleans and processes the raw data, handling missing values and outliers.
  - ❖ Normalizes the data to ensure consistency across different sources.
  - ❖ Extracts relevant features such as access patterns, file modifications, and unusual data transfer activities.

4.3. Threat Intelligence Module:

- **Objective**: Enhance detection capabilities by integrating external threat intelligence.
- **Functionality**:
  - ❖ Aggregates threat intelligence feeds from trusted sources.
  - ❖ Correlates internal data with external threat indicators to identify known malicious activities.
  - ❖ Updates the system with the latest threat signatures and indicators of compromise (IOCs).

4.4. Real-Time Monitoring and Alerting Module:

- **Objective**: Provide continuous surveillance and immediate alerts for potential data leaks.
- **Functionality**:
  - ❖ Monitors data flows and user activities in real-time.
  - ❖ Generates alerts for suspicious activities based on predefined rules and anomaly detection results.

4.5. Incident Response and Mitigation Module:

- **Objective**: Facilitate rapid response to detected data leaks and mitigate their impact.
- **Functionality**:
  - Provides tools for investigating and analyzing alerts.
  - Automates incident response actions such as isolating compromised systems, blocking unauthorized access, and notifying relevant stakeholders.
  - Documents incidents and responses for future reference and compliance reporting.

4.6.Feedback Loop:

- Continuously monitors the performance of the selected nodes.
- Gather feedback to refine prediction models and improve future matchmaking.

5.IMPLEMENTATION:

Implementing a system for "Data Leak Detection and Prevention" involves several steps. Here's a high-level overview of the implementation:

5.1. Understand Requirements

- **Goals:** Detect and prevent data leaks by monitoring data flows, user activities, and system behaviors.

- **Parameters:** Identify key parameters such as data access patterns, file modifications, network traffic, user behavior, and anomalies.
- **Constraints:** Consider constraints like budget, compliance requirements, geographical location, etc.

5.2. Data Collection

- **Historical Data:** Collect historical data on resource usage from various cloud nodes.
- **Real-time Data:** Gather real-time data on current resource availability and performance metrics.

5.3. Data Preprocessing

- **Cleaning:** Clean the data to remove any inconsistencies or missing values.
- **Normalization:** Normalize the data to ensure consistency across different units and scales.
- **Feature Engineering:** Extract and construct relevant features that will be used for prediction.

5.4. Predictive Modeling

- **Model Selection:** Choose appropriate machine learning models (e.g., anomaly detection algorithms, clustering models, classification models) for detecting potential data leaks.

- **Training:** Train the models using historical data to identify patterns indicative of data leaks.

- **Validation:** Validate the models using techniques like cross-validation to ensure accuracy and generalizability.

5.5. Anomaly Detection and Scoring

- **Criteria Definition:** Define criteria for detecting anomalies and potential data leaks based on data access patterns, user behavior, and network activity.
- **Scoring Algorithm:** Develop an algorithm to score each activity based on the defined criteria, indicating the likelihood of a data leak.
- **Ranking:** Set thresholds for triggering alerts based on the anomaly scores.

5.6. Real-time Monitoring and Response

- **Monitoring:** Implement real-time monitoring tools to continuously track data flows, user activities, and system events.
- **Alerting:** Configure alerting mechanisms to notify security teams of potential data leaks based on the anomaly scores and thresholds.
- **Response:** Implement optimization techniques to ensure efficient resource utilization and cost-effectiveness.

5.7. Implementation Framework

- **API Development:** Develop APIs for interacting with the system, allowing users to input security policies, retrieve alerts, and manage responses.
- **Integration:** Integrate the anomaly detection models and monitoring tools into the existing security infrastructure, such as SIEM systems and endpoint protection platform.
- **Monitoring:** Implement monitoring tools to continuously track resource usage and model performance, allowing for real-time adjustments.

5.8. Continuous Improvement

- **Feedback Loop:** Implement a feedback loop to continuously improve the prediction models based on new data.
- **Periodic Review:** Periodically review and update the scoring and matchmaking algorithms to adapt to changing conditions and requirements.
- **User Feedback:** Collect user feedback to improve the system's usability and accuracy.

6.1.CONCLUSION:

Implementing a robust Data Leak Detection and Prevention System is crucial for safeguarding sensitive information. The proposed system employs advanced machine learning techniques, real-time monitoring, and comprehensive data analysis to detect and prevent potential data leaks effectively. Key steps include understanding organizational requirements, collecting and preprocessing data, predictive modeling, real-time monitoring, and automated response. Integration with existing security infrastructure ensures seamless operation, while continuous improvement mechanisms keep the system adaptive to evolving threats.

This comprehensive approach not only protects sensitive data but also ensures regulatory compliance and enhances overall cybersecurity resilience, providing organizations with a reliable defense against data leaks.

6.2.FUTURE SCOPE:

The future scope of the Data Leak Detection and Prevention System includes integrating advanced AI and machine learning techniques for more accurate predictions and anomaly detection. Incorporating blockchain technology can enhance data integrity and traceability. Expanding real-time monitoring capabilities to cover emerging technologies such as IoT and edge computing is crucial. Additionally, enhancing user behavior analytics will provide deeper insights into potential insider threats. Continuous improvement through automated feedback loops and adaptive learning will ensure the system evolves with changing threat landscapes. Collaboration with global threat intelligence networks can further enhance detection capabilities, making the system more robust and comprehensive in protecting sensitive data. The development of user-friendly interfaces and dashboards will facilitate better management and response to potential threats.