JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DATA LEAKAGE DETECTION SYSTEM USING BLOCKCHAIN

Anjali S,
Assistant Professor, Department of computer science
St. Joseph's First Grade College Jayalakshmi Puram Mysore-570012,INDIA

INTRODUCTION

Overview of the Project

The Data Leakage Detection System is a comprehensive solution designed to safeguard sensitive information within an organizational framework. This system offers a range of functionalities catering to both users and administrators. Users can seamlessly register and log in, upload files, control access permissions, and employ watermarking to establish ownership. Robust key authorization and validation mechanisms ensure secure interactions. For administrators, the system provides a dedicated login to access an administrative dashboard. Admins can manage user accounts, overseeing registrations and deletions, upload files for various purposes, and monitor alluploaded files. The system maintains a vigilant stance through comprehensive logging, meticulously tracking user activities, login attempts, file access, and admin actions. The architecture integrates a frontend for user interfaces, a backend handling authentication, access control, and file management, and a database for secure storage of user data, access permissions, and logs. By diligently adhering to role-based access control, encryption, and privacy regulations,the system ensures data integrity and confidentiality. Ultimately, this Data Leakage Detection System fortifies an organization's data governance, empowering users, and administrators to collaborate securely while preempting potential breaches.

Statement of the problem

In today's digital landscape, the proliferation of sensitive data and the rising frequency of data leaks pose significant challenges for organizations. The absence of a robust Data LeakageDetection System leaves these entities vulnerable to breaches that can result in financial losses, reputation damage, and legal repercussions. Existing solutions often lack comprehensive user andadministrator functionalities, failing to address evolving security needs. Furthermore, with users uploading, sharing, and accessing files, there's a pressing need for efficient access control, data tracking, and authorization mechanismsThe potential for insider threats or unauthorized data sharing necessitates an effective system thatnot only enables secure user interactions through key authorization and validation but also empowers administrators to monitor user activities and file access in real time. To address these concerns, our project aims to develop a Data Leakage Detection System that offers a holistic approach to data security. By combining user-friendly features with robust security measures, thissystem will enhance an organization's ability to prevent, detect, and mitigate data leaks, ultimatelyfostering a secure digital environment.

Scope

The scope of this project involves developing a comprehensive Data Leakage Detection System toaddress the escalating challenges of data breaches. The project encompasses user-centric featureslike registration, login, file management, and access control. Administrators gain oversight throughan admin panel. Key-based authorization, real-time monitoring, and logging are integrated to ensure secure interactions and prompt threat detection. The system incorporates encryption, hashing, and watermarking techniques for enhanced security. It will be thoroughly tested for functionality, security, and performance, though external integrations are excluded. The project's focus is on providing a holistic solution that empowers organizations to prevent data leaks, cultivate secure practices, and proactively respond to potential threats.

Motivation

The increasing frequency and severity of data breaches in today's interconnected world underscorethe urgent need for effective data leakage prevention. Our motivation for undertaking this project lies in the desire to provide organizations with a comprehensive and user-centric solution that mitigates the risks associated with data leaks. We recognize the impact of these breaches on businesses, including financial losses, damaged reputation, and compromised customer trust.

By developing a robust Data Leakage Detection System, we aim to empower organizations to proactively safeguard their sensitive information. This system's user-focused features, coupled with stringent security measures, offer a holistic approach to data protection. Through real-time monitoring, access control, and user activity tracking, our project aspires to equip businesses with the tools they need to not only detect potential breaches but also prevent them, ultimatelycontributing to a more secure digital landscape.

Objectives of the study

- Comprehensive Data Protection: Develop a system that ensures the confidentiality and integrity of sensitive data by offering features such as access control, key authorization, and watermarking to prevent unauthorized access and leakage.
- User-Centric Functionality: Design user interfaces that streamline user interactions, allowing seamless registration, login, file management, and access control. Prioritize userexperience to encourage adherence to security protocols.
- Administrator Empowerment: Construct an admin panel that enables administrators to manage user accounts, monitor user activities, and upload files for testing or resource- sharing, bolstering oversight and control.
- Real-Time Monitoring: Implement robust logging mechanisms to track user actions, loginattempts, and file accesses in real time, facilitating prompt detection of suspicious activitiesand potential breaches.
- Key Authorization and Validation: Establish a secure framework for key-based authorization, ensuring that users are granted access only to data and actions they are authorized for.
- Mitigation of Insider Threats: Develop features that allow administrators to identify and respond to insider threats by monitoring unusual user behavior or unauthorized data access.
- Data Leak Prevention: Through proactive measures such as watermarking and access controls, strive to pre-emptively prevent data leaks and unauthorized data sharing.

LITERTAURE REVIEW

Introduction

A literature survey, also known as a literature review, is an essential component of academic research and involves a comprehensive examination and analysis of existing literature on a particular topic. It aims to provide a synthesis of existing knowledge, identify gaps, and highlightareas for further research. In general, a literature survey serves as a foundation for understanding the current state of research in a field, identifying key concepts, theories, methodologies, and trends.

The process of conducting a literature survey involves several steps. First, researchers define the scope and objectives of their study, which helps determine the specific literature to be reviewed. They then conduct a systematic search across various sources, such as academic databases, journals, books, conference proceedings, and relevant websites, to identify relevant literature. Once the literature is gathered, researchers critically evaluate and analyze each source. They assessthe credibility and authority of the authors, evaluate the methodology and quality of the research,and examine the key findings and arguments presented. Researchers also identify common themes, patterns, and gaps in the literature. This analysis helps to develop a coherent and structured overview of the existing knowledge on the topic. In a literature survey, researchers categorize andorganize the reviewed literature based on themes, concepts, or chronological order. This organization aids in synthesizing the information and identifying connections between different studies. Researchers may also use frameworks or models to structure their analysis and provide a comprehensive understanding of the topic. The literature survey provides several benefits to researchers. Firstly, it helps researchers avoid duplicating existing work by identifying what has already been studied and published. It also helps researchers gain a deeper understanding of the research context, theories, and methodologies employed in the field. The survey can highlight controversies, debates, and conflicting findings, leading to a more nuanced understanding of the topic. Furthermore, a literature survey assists in identifying research gaps and unansweredquestions. It reveals areas that require further investigation, allowing researchers to contribute to the existing knowledge base. The survey may also provide insights into potential researchmethodologies and approaches that can be adopted in future studies. In addition to its value in theresearch process, literature surveys are often included as standalone sections in academic papers, theses, and dissertations. These sections provide readers with a comprehensive overview of the existing literature, enabling them to contextualize the new research findings and understand their contribution to the field. To ensure the quality and rigor of a literature survey, it is essential to follow systematic and transparent methods. Researchers should clearly document their search strategies, inclusion and exclusion criteria, and the rationale behind their selection of literature. Itis also important to critically evaluate the strengths and limitations of the reviewed literature and acknowledge any potential biases or gaps in the existing research.

Related Work

1. A System for Detection and Prevention of Data Leak

Technology is growing exponentially in the recent years and most of the organizations store theirdata in digital format. With the rapid growth in technology, there is a need for maintaining securityof data. It is extremely essential as data leak might have a huge effect on the organization. Preventing data leak has become one of the biggest challenges to the organizations. For the securitypurposes, the organizations have implemented several methods like implementation of policies, Firewalls, VPNs, etc. However, with the enhancement of data theft methods, these security measures are not reliable anymore. Hence there was a need for a system that can prevent data leak.Also, as employees have access to sensitive information of the company, they could leak the information either by negligence or on purpose. Hence, securing the data has become a big challenge for the organizations. In this article, we propose a system that will achieve theinformation security goals of the organization, and will be capable of detecting data leak at any state of the data. The proposed system mainly focuses on preventing data leak.

2. Data Leakage detection system

Data Leakage Detection System is one of the most versatile and economical process available forsecuring & Preventing Data. One of the most important key components during the confidential data prevention of these methods is appropriate selection of technological parameters of processing, as well as initial geometrical features of perform, to provide positive functional characteristics of final Data. When any company wants to do any work in minimum time there is one main person who having main server & he passes that data to client server. Client server getsthe data & if required then passes to another if required. In this process i/p address & mac addressis recorded on the main server due to this data leakage from the particular client is found. This client is called as guilt agent.

3. Data Outflow Detection with Guilty Agent Tracking

Data leakage can be defined as accidental leakage of data by a security illiterate or a deliberate attempt by a guilty employee within the firm. When a sender sends confidential data to a third party, for example, a company may have partnerships with other companies that require sharing customer data, there are chances that sensitive data gets leaked accidentally or purposely by an attacker. The research paper focuses on detecting whether the owner's sensitive data has been leaked by trusted third parties, and if possible, identifying the agent that leaked the data. If the unauthorized user is identified, then the owner of the data will get an alert and can stop distributingtheir data with the third parties and send warning alert to that third party that if the same thing happens again by them then the owner of the data can take action against them and even can legallypenalize them for data Cases of leakage. This will further improve the chances of finding the guiltyperson who leaked the data.

4. Data Leakage Detection in Cloud Computing

A sensor network is a group of specialized transducers with a communications infrastructure for monitoring and recording conditions at diverse locations. Cloud computing has been envisioned as the next generation architecture of the IT enterprise due to its long list of unprecedented advantages: on-demand self-service, ubiquitous network access. And This system develops a strengthened security model for considering data security against data leakage attack in cloud andthe storage server in the upload phase of an integrity verification scheme. And presents an efficient verification scheme for ensuring remote data integrity in cloud storage.

Existing System

In the existing system, data leakage detection and prevention are often limited in scope and effectiveness, leaving organizations vulnerable to the escalating threats of data breaches. In this context, manual and disjointed methods are typically employed, lacking the cohesive approach needed to address modern data security challenges. User interactions, such as registration and login, tend to be basic and lack advanced authentication mechanisms. File management and accesscontrol are rudimentary, with limited options for users to define permissions or track data usage. Administrators often lack real-time monitoring capabilities, making it difficult to identify and respond swiftly to suspicious activities. Furthermore, existing systems lack encryption and hashingtechniques, leaving sensitive data susceptible to unauthorized access. The absence of centralizeddashboards for administrators to manage users, files, and system activities hampers oversight andcontrol. Additionally, existing methods do not provide options for watermarking files to establishownership or origin. In essence, the existing system fails to meet the dynamic data security requirements of the digital age. Its limitations contribute to an environment where data breaches are more likely to occur, potentially resulting in severe consequences for organizations. As technology evolves and threats become more sophisticated, the shortcomings of the existing system highlight the pressing need for a comprehensive and advanced Data Leakage Detection System that effectively safeguards sensitive information and mitigates the risks of data leaks.

Drawbacks

- False Positives: Data leakage detection systems often generate false positive alerts, flagging normal or authorized data transfers as potential leaks. This can lead to unnecessaryinvestigations and wasted resources.
- False Negatives: On the flip side, these systems can also miss actual data leaks, failing to detect breaches when they occur. This can result in sensitive data being exposed without detection.
- Complex Configuration: Many data leakage detection systems require complex configurations and fine-tuning to work effectively. This can be time-consuming andchallenging for organizations with limited resources or expertise.
- Scalability: Some systems may struggle to scale effectively as an organization's data volume grows. This can lead to performance issues and decreased accuracy.
- Privacy Concerns: In order to monitor and detect data leakage, these systems often requireaccess to sensitive data and may raise privacy concerns among employees and stakeholders.
- Evolving Threats: Data leakage detection systems may struggle to keep up with evolving and sophisticated data leakage techniques used by attackers. New attack vectors and evasion tactics can bypass these systems.
- Resource Intensive: Running data leakage detection systems can consume significant computational resources, which may lead to performance issues and increased operationalcosts.
- Limited Coverage: Some systems may focus on specific data channels (e.g., email, web traffic) and may not provide comprehensive coverage across all data transfer methods andcommunication channels.

- False Alarms: Alerts generated by these systems can sometimes overwhelm security teamswith a high volume of false alarms, leading to alert fatigue and reduced responsiveness toreal threats.
- Integration Challenges: Integrating data leakage detection systems with existing security infrastructure and policies can be challenging and may require custom development and ongoing maintenance.
- Lack of User Awareness: Users may not be aware of the system's presence or its rules, which can result in unintentional violations and data leakage.
- Cost: Implementing and maintaining data leakage detection systems can be expensive, making them less accessible to smaller organizations with limited budgets.
- Limited Visibility: Some systems may lack visibility into encrypted traffic, making it difficult to detect data leaks in encrypted communication channels.
- Compliance Challenges: While these systems can help with compliance requirements, theymay not provide all the necessary features to meet specific regulatory standards, requiringadditional tools and processes.
- Dependence on Signatures: Some data leakage detection systems rely heavily on signature-based detection, making them less effective against zero-day threats and new attack patterns.
- To mitigate these drawbacks, organizations should carefully evaluate and select data leakage detection solutions that align with their specific needs and capabilities, andcomplement these systems with a holistic cybersecurity strategy.

Proposed work

The proposed system constitutes an advanced Data Leakage Detection System designed toholistically address the intricate challenges of data security and leak prevention. The system is characterized by its user-centric features, robust security measures, and real-time monitoring capabilities. At its core, the system offers a user-friendly interface, facilitating seamless user registration and blocks creation, login and blocks creation, and file management. Users can exercise fine-grained control over file access permissions, with key-based authorization ensuring secure interactions. Furthermore, users have the option to watermark uploaded files, enabling effective tracking of data ownership and origin. For administrators, the system presents acomprehensive admin panel. This panel empowers administrators to manage user accounts, uploadfiles using block generation for various purposes, and monitor user activities and system logs. Thereal-time monitoring and logging system provides insights into user actions, login attempts, and file accesses, facilitating rapid detection of suspicious activities. The proposed system operates within a robust architectural framework, incorporating encryption mechanisms to protect data at rest and during transmission. Stringent hashing techniques are employed to safeguard password storage. Rigorous testing and validation protocols ensure that the system functions flawlessly, meeting security and performance benchmarks. Overall, the proposed Data Leakage Detection System aspires to equip organizations with an effective toolset to mitigate the risks of data breaches, fortify their data security strategies, and foster a secure digital environment conducive to collaboration and innovation.

SOFTWARE REQUIREMENTS SPECIFICATION

Introduction

The Software Requirements Specification (SRS) serves as a foundational document that outlines the comprehensive details and expectations of the proposed Data Leakage Detection System. Thisdocument serves as a roadmap that communicates the project's scope, functionalities, constraints,and design considerations to all stakeholders involved. In light of the increasing prevalence of data breaches and the critical importance of data security in today's digital landscape, the SRS is designed to address the limitations of existing systems. The proposed Data Leakage Detection System aims to provide a holistic and robust solution that encompasses user-centric features and stringent security measures. The introduction section of the SRS provides an overview of the project's goals, emphasizing the need for a more effective approach to data leak prevention and detection. It highlights the limitations of traditional methods and sets the stage for the advanced features and functionalities that the proposed system will offer. By presenting the motivation andobjectives, this section lays the foundation for the subsequent detailed specifications that will guidethe development, testing, and deployment phases of the project. Ultimately, the SRS introduction contextualizes the project within the broader landscape of data security, underscoring its significance and relevance in safeguarding sensitive information.

Project Functions

 User Registration and Authentication: Users can create accounts securely through a registration process. The system employs robust authentication mechanisms, including multi-factor authentication, to ensure authorized access with blocks generation.

 User and Admin Login: Blocks are created for both users and administrators to securelylog into the system using their respective credentials, allowing access to their designated functionalities.

 File Upload and Management: Blocks are generated for ledger management so users canupload files securely to the system, managing them by setting access permissions, choosingwatermark options, and categorizing them for easy retrieval.

 Access Control: The system offers granular control over who can access uploaded files. Users can define permissions and restrict access based on user roles.

 Key Authorization and Validation: Users employ access keys or tokens to validate theirauthorization for specific actions. The system verifies these keys before allowing file downloads or access to certain functionalities.

 Real-Time Monitoring and Logging: The system continuously monitors user activities, capturing events like file accesses, logins, and user interactions. Real-time logging aids inthe prompt detection of suspicious behavior.

 Admin Oversight: Administrators can view and manage user accounts, upload files for testing or resource-sharing, and oversee system activities through a dedicated admin panel.

 File Download: Authorized users can securely download files they have access to,ensuring data confidentiality and blocks mapping is done.

 Watermarking: Users can choose to apply watermarks to their uploaded files, providing a traceable connection to the source and ownership of the data.

 Security Measures: The system integrates encryption techniques to protect data at rest and in transit. Passwords are stored securely using hashing algorithms.

 Comprehensive Reporting: The system generates comprehensive reports, including useractivities, access logs, and potential security incidents, aiding in compliance and risk management.

Hardware and Software Requirements

Hardware Requirements

- Processor : Intel i3 2.53GHz
- Hard Disk : 40GB
- RAM : 8 GB or above

Software Requirements

- Operating system : Windows 8 and above
- Front End : HTML, CSS, JS.
- Coding Language : Python.
- IDE : Visual Studio Code

Functional Requirement

 User Registration and Login

- Users can create accounts with unique usernames and passwords.
- Users can log in securely using their credentials.

 File Upload and Management

- Users can upload files to the system.
- Users can categorize, edit metadata, and manage their uploaded files.

 Access Control

- Users can set access permissions for their files, defining who can view and download them.
- Administrators can modify access permissions if necessary.

 Key Authorization and Validation

- Users need valid access keys to perform certain actions.
- The system validates access keys before allowing authorized actions.

 Real-Time Monitoring and Logging

- The system records user activities, login attempts, and file accesses in real time.
- Administrators can view logs for auditing and security purposes.

 Administrator Functionality

- Administrators can manage user accounts and permissions.
- Administrators can upload files for various purposes.
- Administrators can view user activities and system logs.

 File Download

- Authorized users can securely download files they have access to.

 Watermarking

- Users can choose to apply watermarks to their uploaded files for tracking ownership.

Non-Functional Requirement

 Security

- Data is encrypted during transmission and storage.
- Passwords are securely hashed and stored.
- Access control mechanisms prevent unauthorized access.

- **Performance**

- The system should handle concurrent user interactions without significant delays.
- File uploads and downloads should be efficient even with large file sizes.

- **Usability**

- User interfaces should be intuitive and user-friendly.
- Admin panel should be organized and easy to navigate.

- **Reliability**

- The system should be available and responsive most of the time.
- Backup and recovery mechanisms should be in place to minimize data loss.

- **Scalability**

- The system should be designed to handle an increasing number of users and files.

- **Compliance**

- The system should adhere to data protection and privacy regulations.

- **Reporting**

- The system should generate comprehensive reports on user activities and system logs.

- **Integration**

- The system should integrate with existing authentication systems if required.

- **Documentation**

- Comprehensive documentation should be provided for users, administrators, anddevelopers.

User Requirements

- **User Registration and Login**

- Users should be able to create accounts easily with unique usernames and secure passwords.

- A password recovery mechanism should be in place for forgotten passwords.
- Logging in should be straightforward, with the option for multi-factor authentication foradded security.

 File Upload and Management

- Users need to easily upload files of various formats and sizes.
- They should be able to view, edit metadata, and organize their uploaded files in a user-friendly interface.

 Access Control

- Users should be able to specify who can access their uploaded files by setting permissionlevels.
- Flexibility in defining read and download permissions for specific users or groups is desired.

 Key Authorization and Validation

- Users should be able to generate, manage, and use access keys for secure interactions.
- The system should notify users about key validation status.

 Real-Time Monitoring and Logging

- Users should be able to view their own activity logs and login history for added securityawareness.

 File Download

- Authorized users should be able to securely download files they have access to, without anytechnical complexities.

 Watermarking

- Users should have the option to apply watermarks to their uploaded files, with customizablewatermark content.

Constraints

- Budget and Resources: The project must adhere to budget constraints, including expensesrelated to server hosting, app development, and potential third-party services.
- Time Constraints: The project timeline should consider development, testing, and potentialdelays to ensure timely completion.
- Data Privacy Regulations: The system must adhere to data protection regulations andensure user data privacy.
- NLP Complexity: The complexity of implementing NLP techniques might requirespecialized skills and resources for accurate query processing.
- Security Measures: Implementing strong security measures to prevent data breaches andunauthorized access is crucial.
- Server Reliability: The Flask server hosting the system must maintain high uptime andreliability to ensure

uninterrupted service.

Feasibility study

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is nota burden to the company.

Three key considerations involved in the feasibility analysis are:

- **Economical Feasibility**

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development ofthe system is limited. The expenditures must be justified. Thus the developed system as well withinthe budget and this was achieved because most of the technologies used are freely available. Onlythe customized products had to be purchased.

- **Technical Feasibility**

This study is carried out to check the technical feasibility, that is, the technical requirementsof the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement,as only minimal or null changes are required for implementing this system.

- **Social Feasibility**

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the userssolely depends on the methods that are employed to educate the user about the system and to makehim familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

- **Schedule Feasibility**

The time schedule required for the development of this project is very important since over-runs result in escalated projects costs and also hinders in the development of the other systems.

Modules

- **User Management:** This module encompasses user registration, authentication, andprofile management. It allows users to create accounts, securely log in, and manage their profile details.

- **Query Processing:** The heart of the system, this module utilizes NLP techniques for queryanalysis. It recognizes user intent, extracts keywords, and processes queries related to college information.
- **Response Generation:** Based on processed queries, this module generates contextually relevant responses containing accurate details about courses, faculties, and facilities.
- **Android Application Interface:** This module involves the development of the Android app interface, including user registration, login, and query submission functionalities.
- **API Communication:** Responsible for establishing API endpoints, this module enables communication between the Android app and the Flask server for query processing and response retrieval.
- **Database Management:** This module manages the secure storage of user profiles, query history, and college data. It ensures data integrity, security, and efficient retrieval.

SYSTEM DESIGN

Introduction

The system design is a pivotal phase in the development of the Data Leakage Detection System, serving as the blueprint for translating user requirements into a tangible and functional solution. This phase involves conceptualizing the architecture, components, and interactions that will collectively constitute the envisioned system. The introduction to the system design elaborates onthe project's goals and the imperative for an advanced data leakage prevention solution. It underscores the escalating threats of data breaches in the digital landscape and emphasizes the need for a robust system that aligns with user requirements and industry best practices. The introduction also highlights the significance of a well-structured system design in ensuring the system's effectiveness, security, and scalability. By providing an overview of the design process and its key objectives, stakeholders gain insight into how the project will address the challenges and complexities of data security and leak prevention. In essence, the introduction to the system design lays the foundation for the subsequent sections, setting the context for the architectural decisions, component interactions, and security measures that will be elaborated upon. It serves asa prelude to the detailed design considerations that will guide the development and implementation phases of the Data Leakage Detection System.

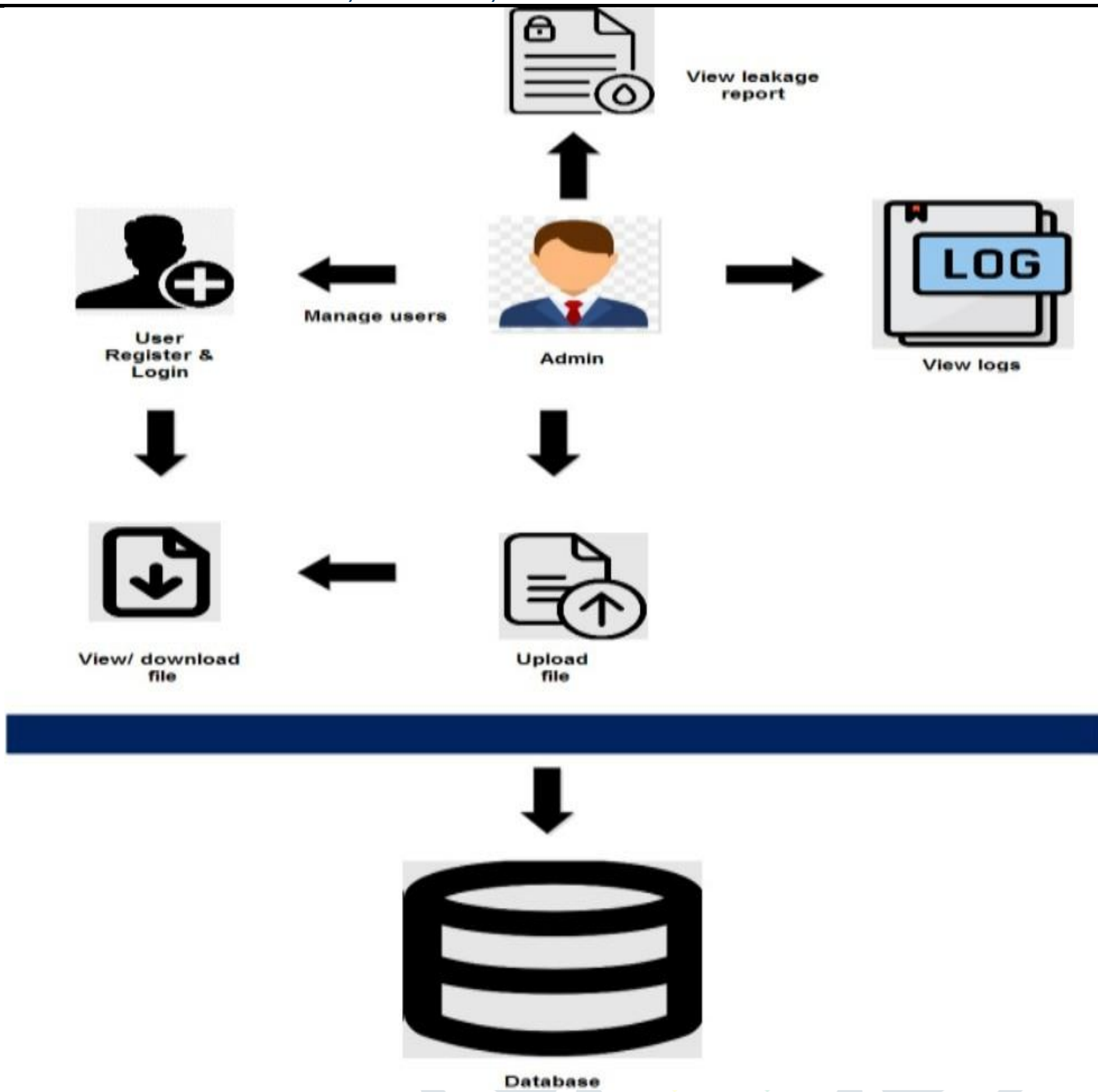

Figure 4.1: System Architecture

The system architecture of the Data Leakage Detection System is a carefully designed structure that outlines the arrangement and interactions of its components to ensure the effective functioningof the system. It is devised to comprehensively address the complex challenges associated with data security and leakage prevention. At its core, the architecture comprises several key components, each serving a specific purpose. The user interfaces provide users and administratorswith intuitive interaction points. The authentication and authorization mechanisms ensure secure

access to the system's functionalities, while the file management module empowers users to upload, categorize, and manage files securely. The real-time monitoring and logging module captures and records user activities and system events for timely threat detection. Central to the architecture is the access control and key validation system, which guarantees that data is accessedonly by authorized individuals using valid keys. The security and encryption layer employs advanced techniques to safeguard data integrity and confidentiality. The database stores user information, access permissions, and logs securely. This architecture facilitates a seamless flow of interactions, starting from user interfaces and passing through various modules to ensure securefile management, access control, and real-time monitoring. By emphasizing security, efficiency, and user-friendliness, the system architecture provides a comprehensive solution for organizationsto proactively protect their data and prevent leakage.

Detailed design

Detailed design starts after the system design phase is completed and the system design has been certified through the review. The goal of this phase is to develop the internal logic of each of the modules identified during system design.

In the system design, the focus is on identifying the modules, whereas during detailed design the focus is on designing the logic for the modules. In other words in system design attention is on what components are needed, while in detailed design how the components can be implemented in the software is the issue.

The design activity is often divided into two separate phase system design and detailed design. System design is also called top-level design. At the first level focus is on deciding which modules are needed for the system, the specifications of these modules and how the modules should be interconnected. This is called system design or top-level design. In the second level the internal design of the modules or how the specifications of the module can be satisfied is decided. This design level is often called detailed design or logic design.

Data Flow Diagram

DFD graphically representing the functions, or processes, which capture, manipulate, store, and distribute data between a system and its environment and between components of a system. The visual representation makes it a good communication tool between User and System designer. Structure of DFD allows starting from a broad overview and expand it to a hierarchyof detailed diagrams. DFD has often been used due to the following reasons:

- Logical information flow of the system
- Determination of physical system construction requirements
- Simplicity of notation
- Establishment of manual and automated systems requirements

Basic Notation:

Figure 4.2: Data Flow Diagram

Process: any process that changes the data, producing an output. It might perform computations, or sort data based on logic, or direct the data flow based on business rules. A short label is used todescribe the process, such as "Submit payment."

Data store: files or repositories that hold information for later use, such as a database table or a membership form. Each data store receives a simple label, such as "Orders."

External entity: an outside system that sends or receives data, communicating with the system being diagrammed. They are the sources and destinations of information entering or leaving the system. They might be an outside organization or person, a computer system or a business system.They are also known as terminators, sources and sinks or actors. They are typically drawn on the edges of the diagram

Data flow: the route that data takes between the external entities, processes and data stores. It portrays the interface between the other components and is shown with arrows, typically labeled with a short data name, like "Billing details.

Data Flow Diagram

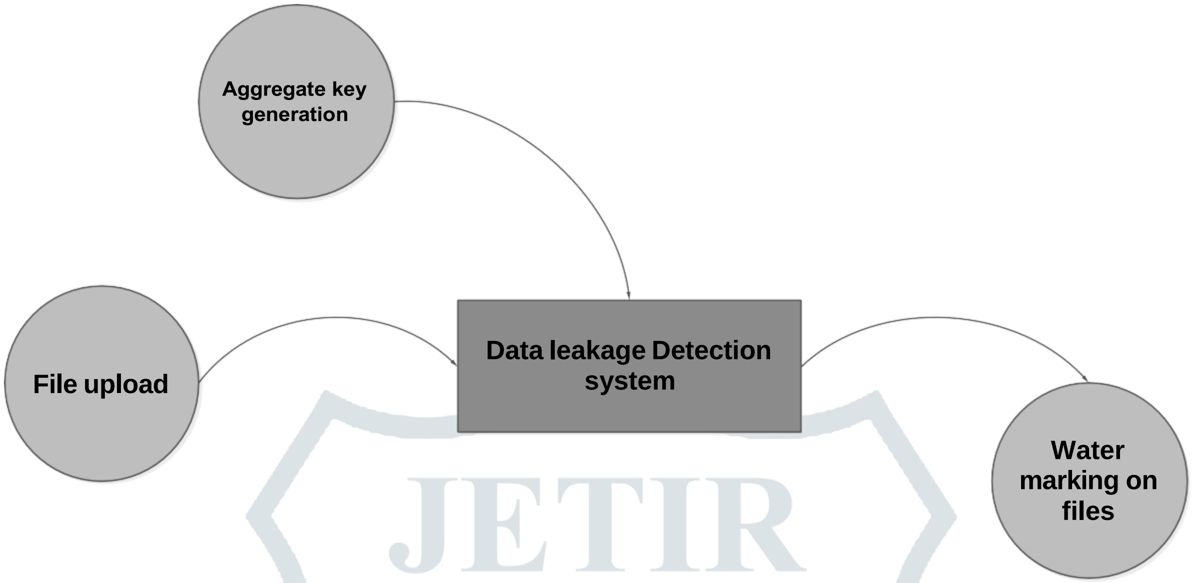

Figure 4.3: Level 0 DFD

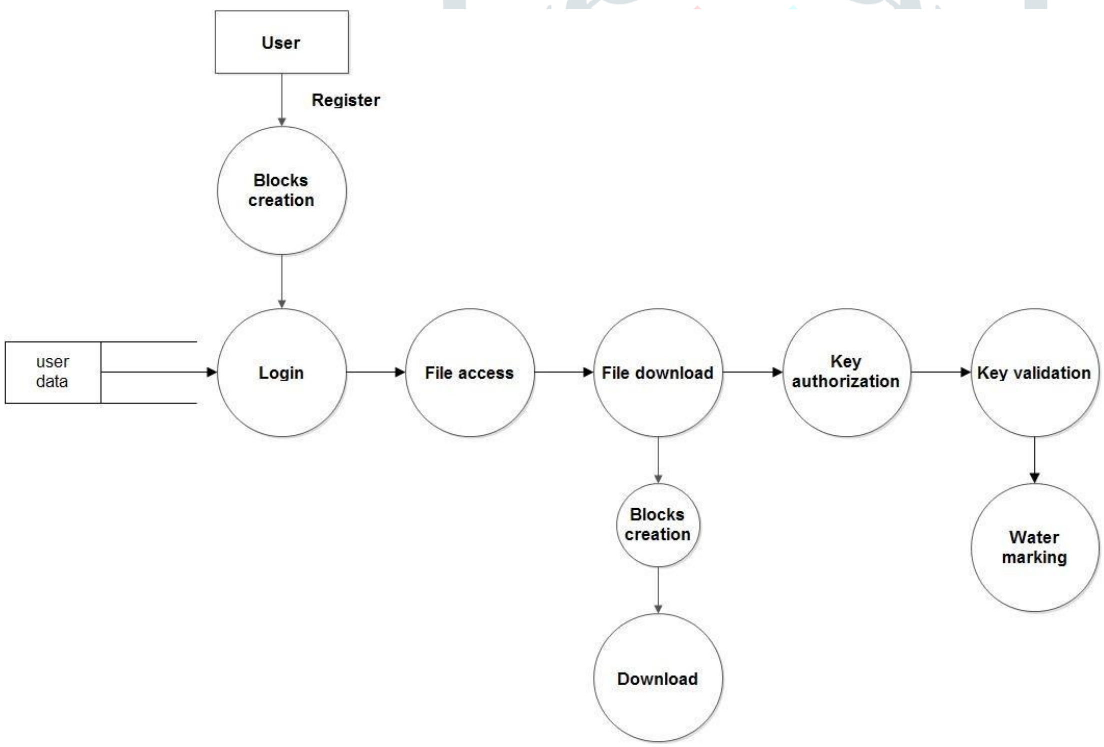

Figure 4.4: User DFD

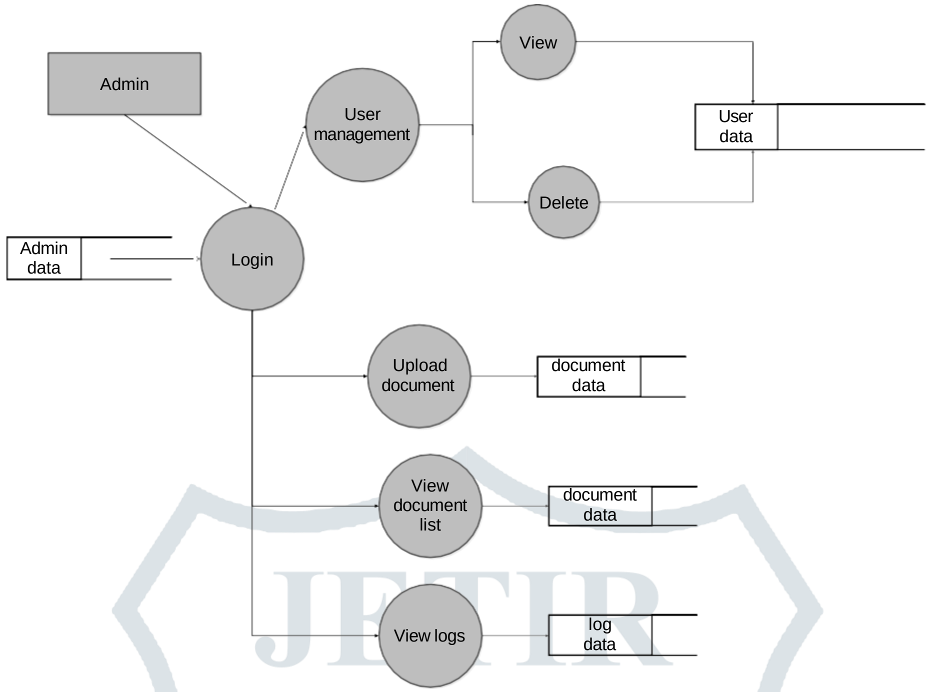

Figure 4.5: Admin DFD

Use case diagram

Use case diagram is a graph of actors, a set of use cases enclosed by a system boundary, communication associations between the actor and the use case. The use case diagram describes how a system interacts with outside actors; each use case represents a piece of functionality that a system provides to its users. A use case is known as an ellipse containing the name of the use case and an actor is shown as a stick figure with the name of the actor below the figure.

The use cases are used during the analysis phase of a project to identify and partition system functionality.They separate the system into actors and use case. Actors represent roles that are played by user of the system. Those users can be humans, other computers, pieces of hardware, or even other software systems.

Figure 4.6: User use case Diagram

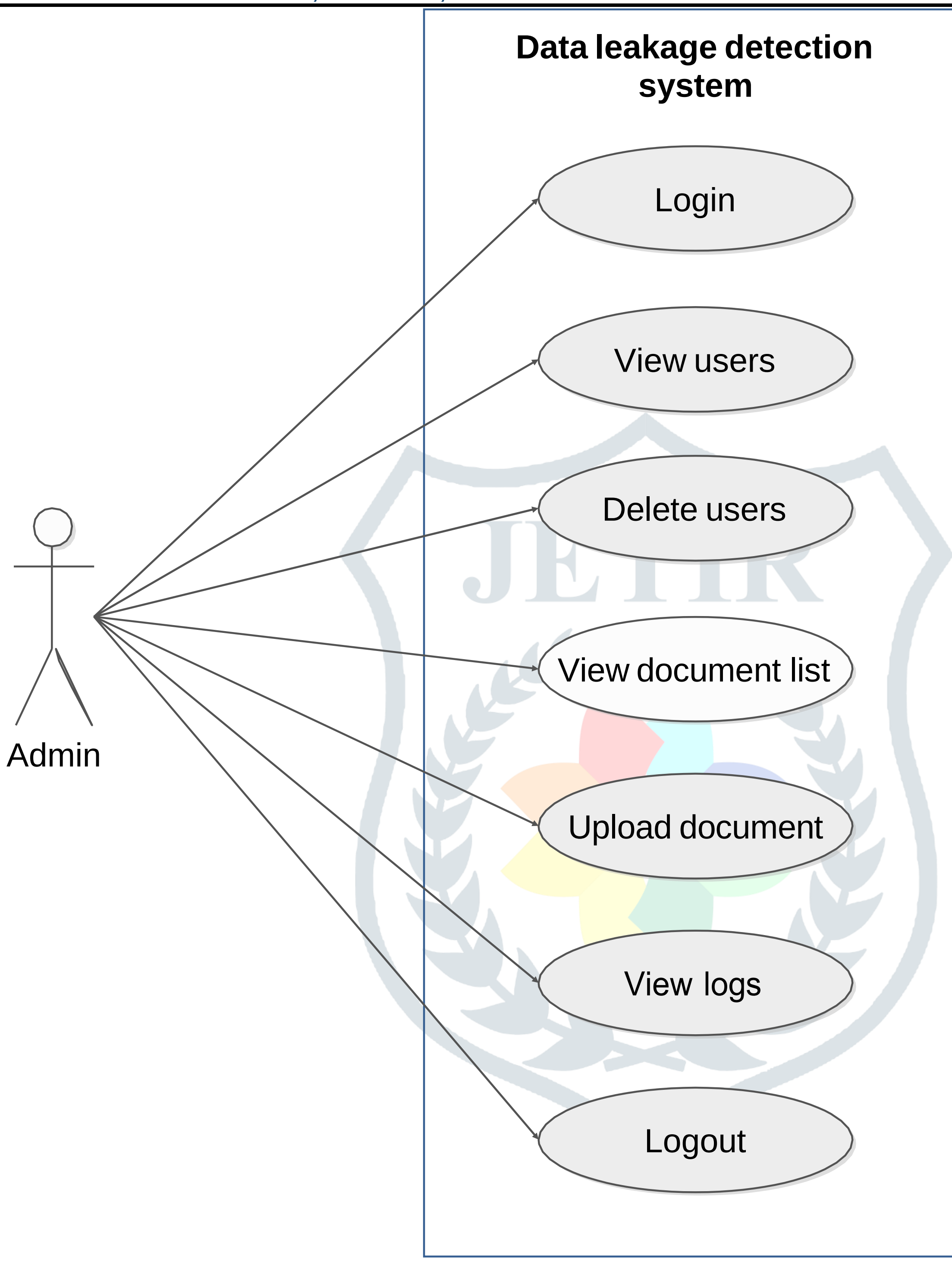

Figure 4.7: Admin use case Diagram

Sequence Diagram

A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classesinvolved in the scenario and the sequence of messages exchanged between the objects needed to carry outthe functionality of the scenario. Sequence diagrams are sometimes called **event diagrams**, **event scenarios.**

UML sequence diagrams are used to represent or model the flow of messages, events and actions betweenthe objects or components of a system. Time is represented in the vertical direction showing the sequence of interactions of the header elements, which are displayed horizontally at the top of the diagram SequenceDiagrams are used primarily to design, document and validate the architecture, interfaces and logic of the system by describing the sequence of actions that need to be performed to complete a task or scenario. UML sequence diagrams are useful design tools because they provide a dynamic view of the system behavior.

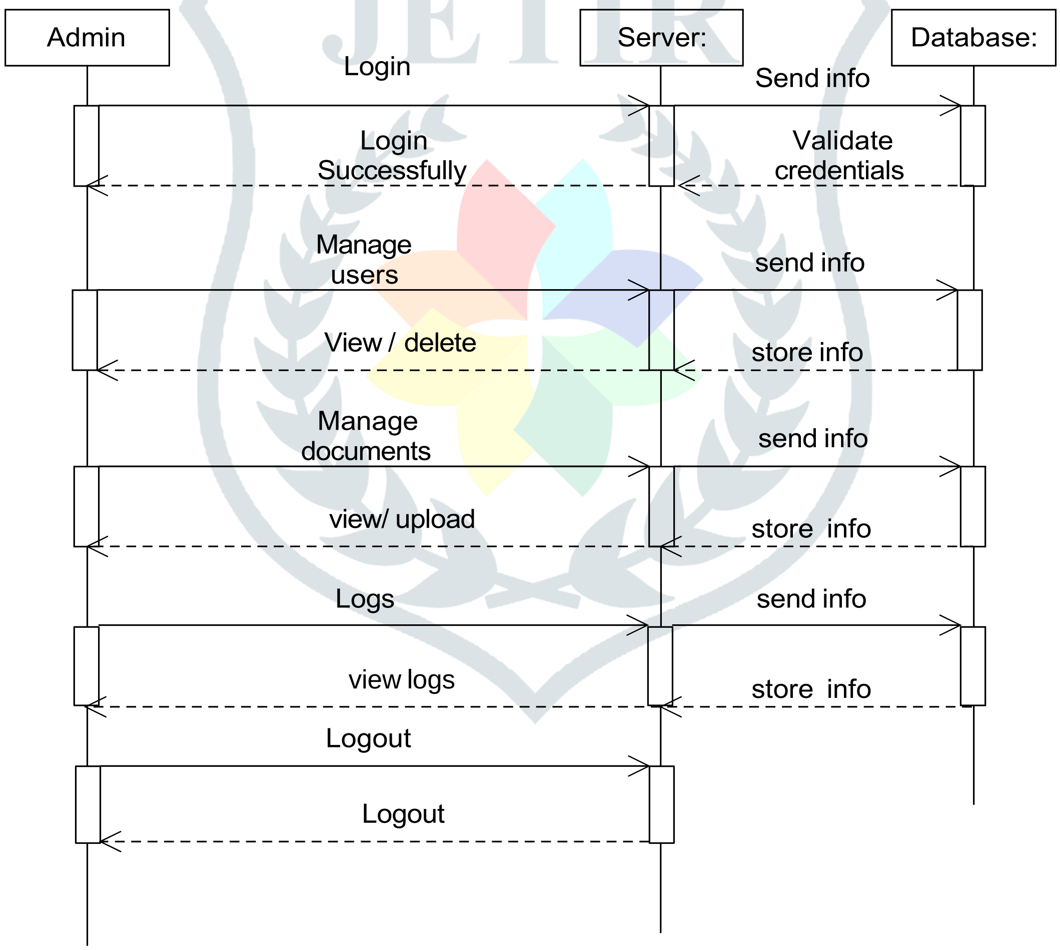

Figure 4.8: Admin sequence diagram

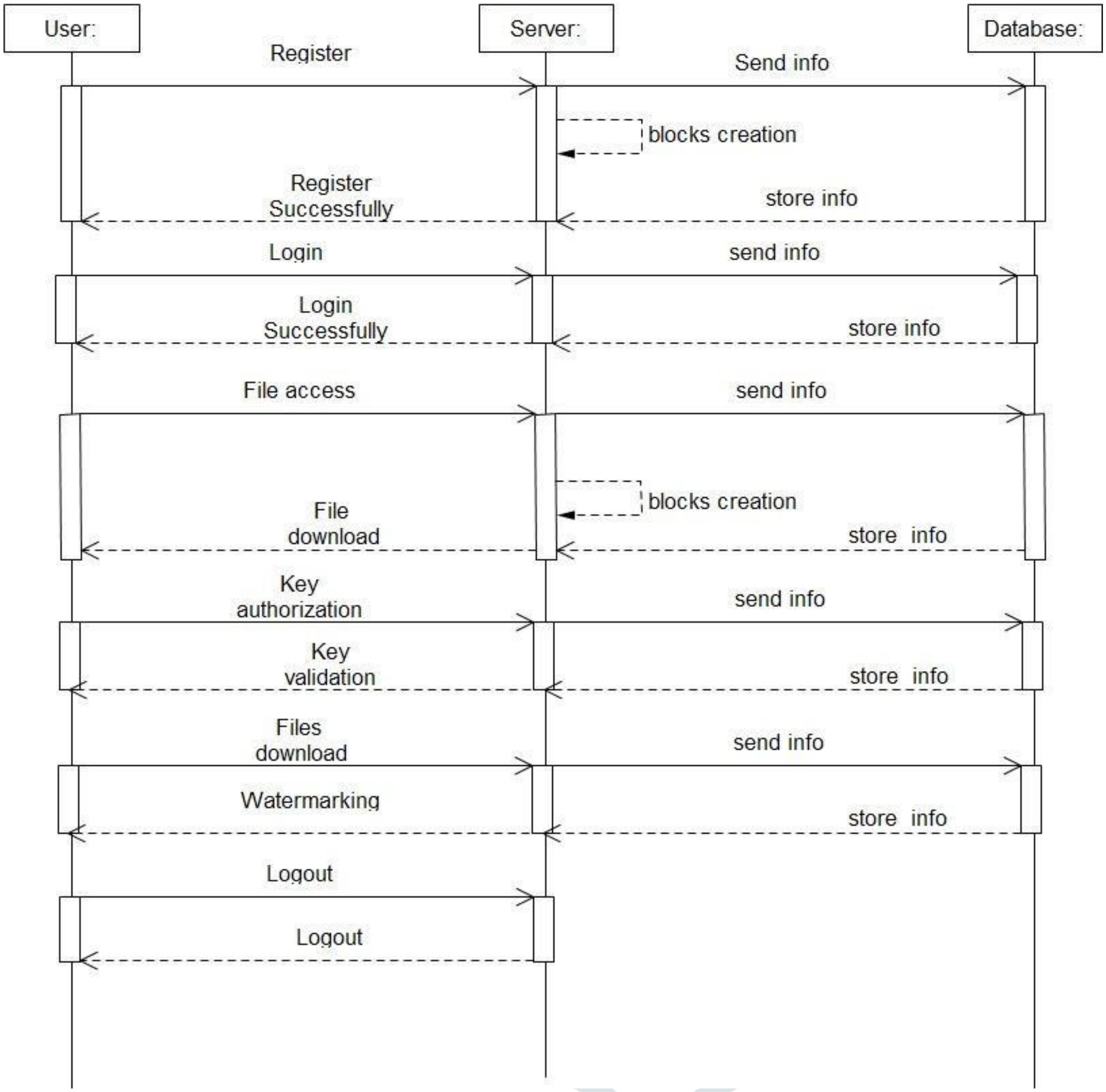

Figure 4.9: User sequence Diagram

Activity Diagrams

Activity diagrams represent the business and operational workflows of a system. An activity diagram is a dynamic diagram that shows the activity and the event that causes the objectto be in the particular state. It is a simple and intuitive illustration of what happens in a workflow,what activities can be done in parallel, and whether there are alternative paths through the workflow.

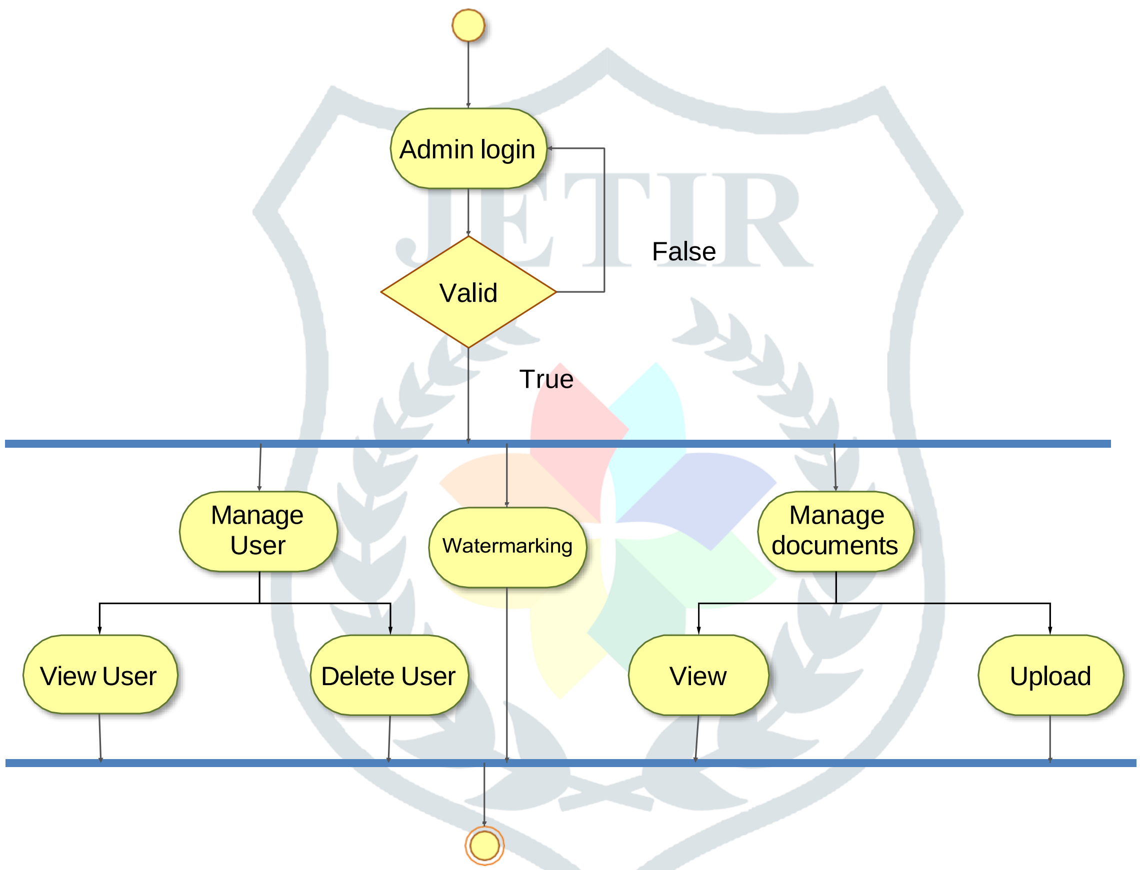

Figure 4.10: Admin Activity

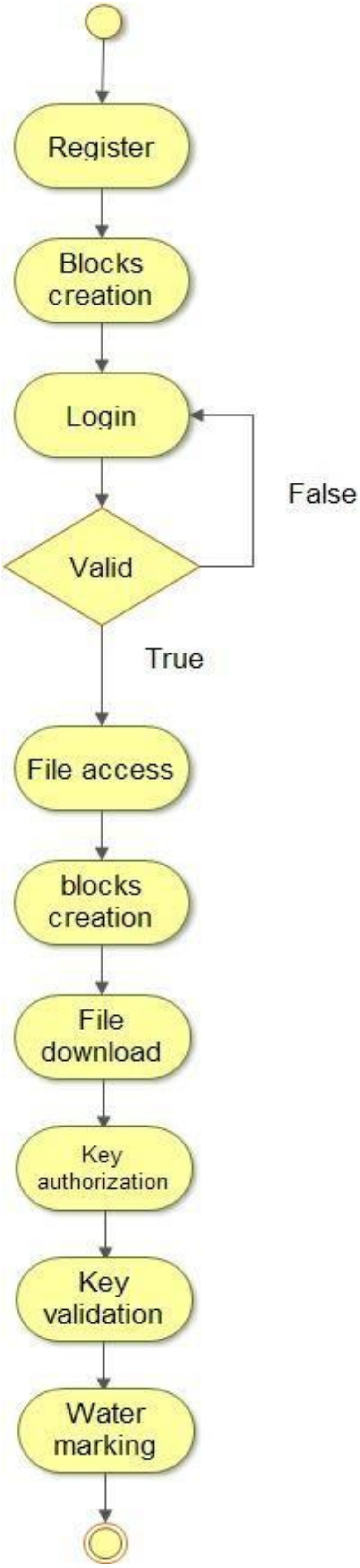

Figure 4.11: User Activity

Entity-Relationship Diagram

Entity Relationship Diagram depicts the various relationships among entities, considering each objective as entity. Entity relationships are described by their dependence on each other, as well as the extent of the relationship between the data stores. It depicts the relationship between data objects. The ER diagram is a notation that is used to conduct the data modeling activity.

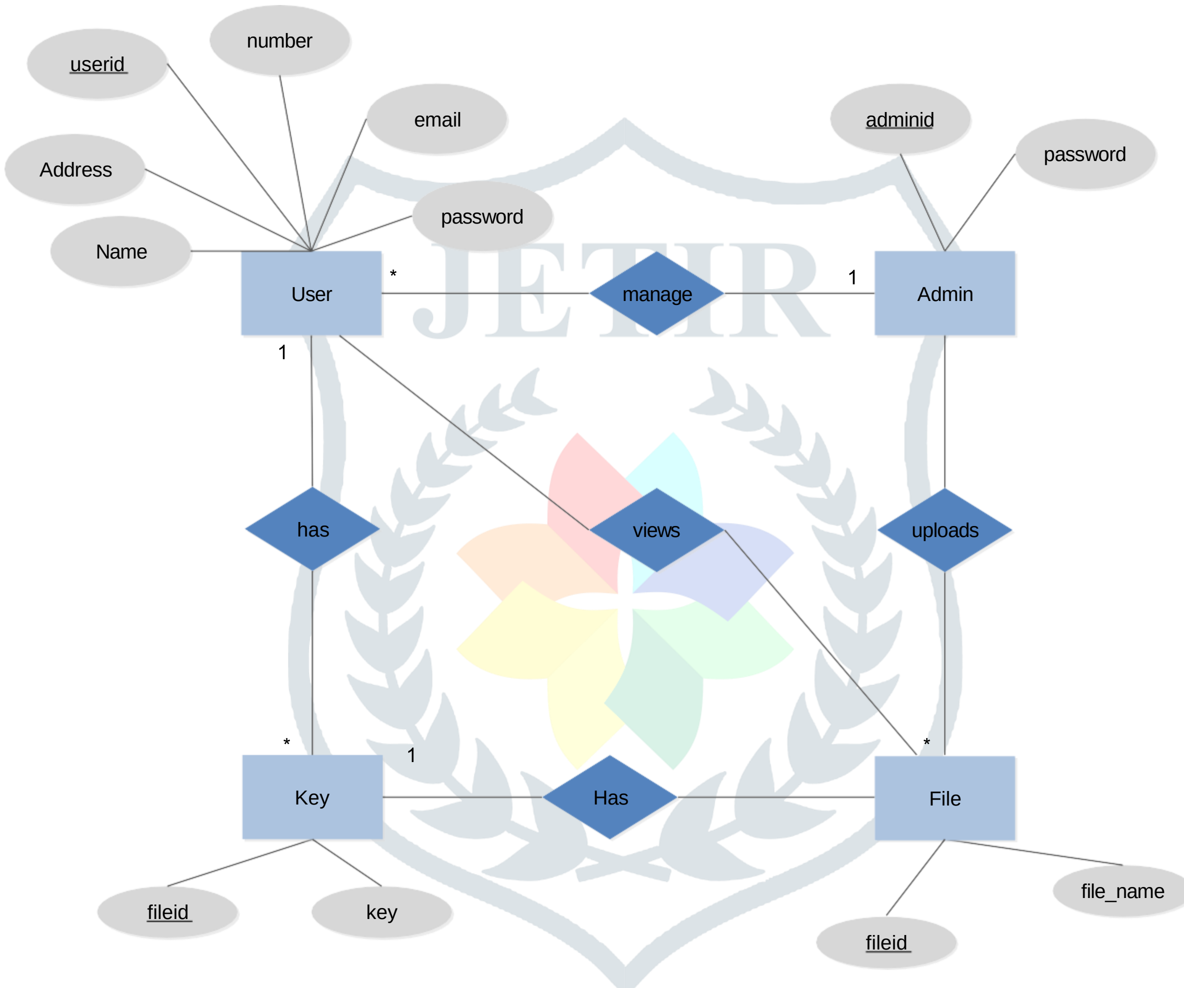

Figure 4.12: Entity-Relationship Diagram

SYSTEM IMPLIMENTATION

Introduction

The system implementation phase marks a critical juncture in the development of the Data LeakageDetection System, where the meticulously designed system architecture transitions from concept to reality. This phase involves translating the architectural blueprints and design specifications intofunctional code, modules, and components that collectively bring the envisioned system to life. The introduction to the system implementation delineates the significance of this phase within thebroader context of the project. It highlights how the careful planning, meticulous design, and thoughtful consideration of user requirements culminate in the development of a tangible solutionthat addresses the pressing challenges of data security and leak prevention. Furthermore, the introduction underscores the importance of precision and attention to detail in coding, integratingcomponents, and validating functionalities. It emphasizes the need for adherence to best practices,industry standards, and the security measures outlined in the system architecture. By providing anoverview of the system implementation phase, the introduction sets the stage for the subsequent sections that will delve into the technical specifics of coding, integration, testing, and quality assurance. It serves as a prelude to the detailed descriptions that guide developers in transformingthe planned system into a fully functional reality, ready to fortify organizations' data security strategies.

Algorithms Used

Blockchain

Blockchain is a distributed and decentralized digital ledger technology that records transactions across multiple computers in a way that ensures the security, transparency, and immutability of the data. It was originally created as the underlying technology for the cryptocurrency Bitcoin buthas since found applications in various other fields beyond cryptocurrencies.

Here are some key characteristics and concepts associated with blockchain:

- Decentralization: Unlike traditional centralized systems (like banks or centralized databases),blockchain operates on a network of computers (nodes) that are distributed across the globe. Each node has a copy of the entire blockchain, and no single entity or authority has control over the network.
- Transparency: Transactions recorded on a blockchain are visible to all participants in the network. This transparency helps to establish trust among users, as they can independently verify transactions.
- Security: Blockchain uses cryptographic techniques to secure transactions and control accessto the data. Once a transaction is added to the blockchain, it becomes very difficult to alter ordelete it, making the system highly secure against fraud and tampering.
- Immutability: The data recorded on a blockchain is considered immutable, meaning it cannotbe changed or deleted. This property is crucial for maintaining the integrity of transaction history.
- Consensus Mechanisms: Blockchain networks use consensus mechanisms, such as Proof of Work (PoW) or Proof of Stake (PoS), to validate and agree on the order of transactions. Thesemechanisms ensure that all nodes in the network reach a consensus on the state of the blockchain.
- Smart Contracts: Blockchain platforms like Ethereum support smart contracts, which are self-executing contracts with predefined rules and conditions. Smart contracts automatically execute when their conditions are met, eliminating the need for intermediaries in many contractual agreements.
- Public and Private Blockchains: There are public blockchains, like Bitcoin and Ethereum, thatare open to anyone and are maintained by a decentralized community. There are also private blockchains, which restrict access and are often used by organizations for internal purposes.
- Cryptocurrencies: Blockchain technology gave rise to cryptocurrencies like Bitcoin, which are digital assets that use blockchain for secure and transparent transactions. However, blockchain has applications beyond

cryptocurrencies, including supply chain management, voting systems, healthcare records, and more.

- Mining: In PoW-based blockchains like Bitcoin, miners use computational power to solve complex mathematical puzzles. The first one to solve the puzzle gets the right to add a new block of transactions to the blockchain and is rewarded with cryptocurrency.
- Blockchain technology has the potential to revolutionize various industries by providing a secure and efficient way to record and verify transactions, reduce fraud, and streamline processes. Its applications continue to expand as more people explore its capabilities.

Proof of Work

PoW stands for "Proof of Work," and it is a consensus algorithm used in blockchain technologyto secure and validate transactions on a decentralized network. Here's how it works:

- **Mining:** In a PoW-based blockchain network, participants known as miners compete to solve complex mathematical puzzles. These puzzles are called "hash functions."
- **Difficulty:** The network adjusts the difficulty of these puzzles regularly to ensure that a new block is added to the blockchain at a relatively constant rate, typically every 10 minutes in the case of Bitcoin.
- **Computation:** Miners use their computational power (usually in the form of specialized hardware called ASICs or Graphics Processing Units) to attempt to find a solution to the puzzle. They do this by repeatedly hashing the contents of the new block (including a reference to the previous block) with a nonce (a random number).
- **Proof of Solution:** When a miner finds a solution that meets the current difficulty level (i.e., the hash value of the block is below a certain target value), they broadcast this solutionto the network.
- **Consensus:** Other nodes on the network can easily verify that the miner's solution is correctby checking the hash value. Once confirmed, the new block is added to the blockchain, andthe miner is rewarded with newly created cryptocurrency (e.g., Bitcoin) and transaction fees.
- **Security:** PoW is considered a secure consensus mechanism because it requires a significant amount of computational power to alter the blockchain's history. To manipulatethe blockchain, an attacker would need to control more than 50% of the network's total computational power, which is known as a "51% attack." This is economically infeasible for most blockchains with a substantial user base.

While PoW is known for its security, it has drawbacks, including high energy consumption due to the computational work required for mining. In response to these concerns, alternative consensus algorithms like Proof of Stake (PoS) have been developed, which aim to provide similarsecurity with lower energy consumption.

Experimental Results

6.1 Screenshots

Figure 6.1: login page **Figure 6.2: Upload Page**

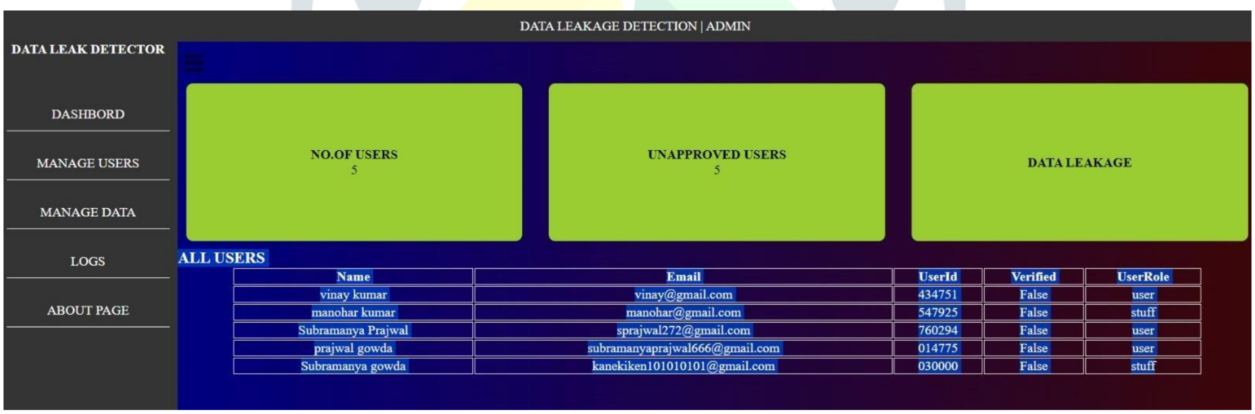

Figure 6.3: Admin Page

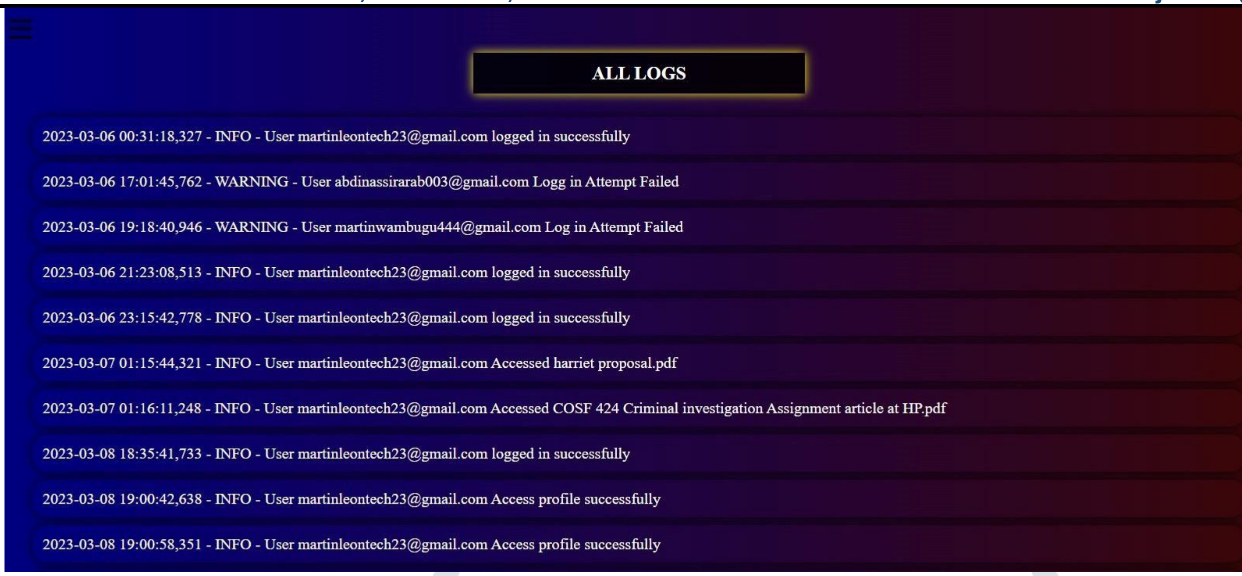

Figure 6.4: Logs Page

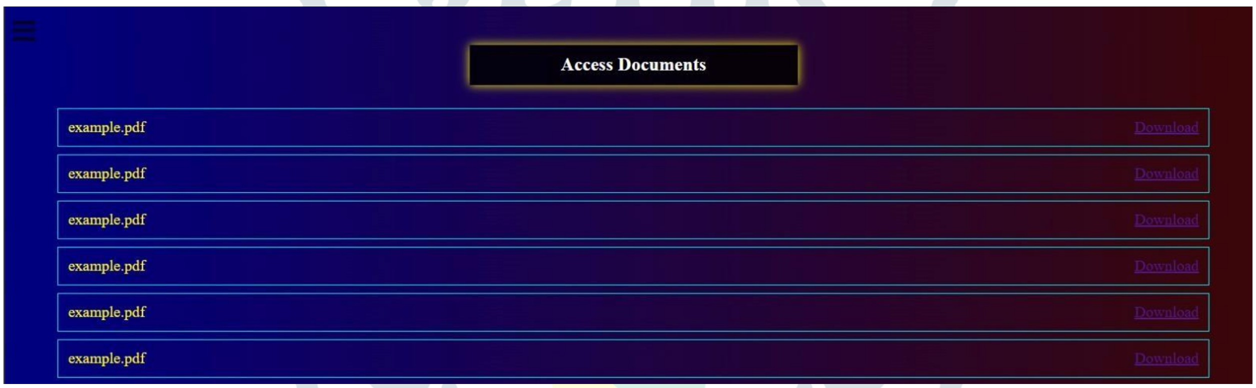

Figure 6.5: Document Access Page

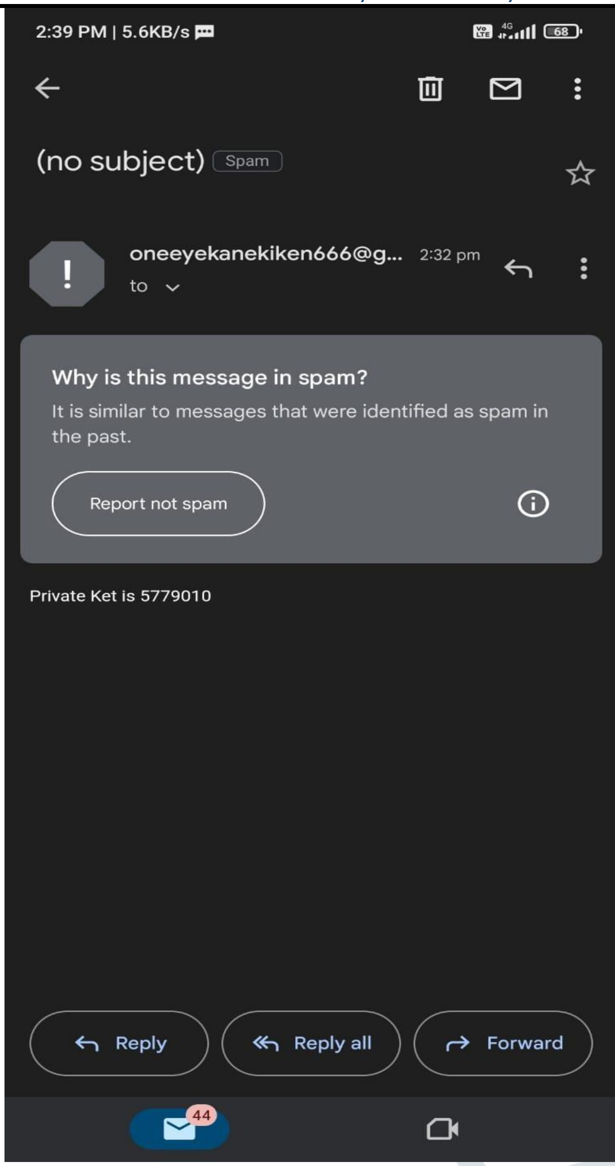

Figure 6.6: Private Key Mail

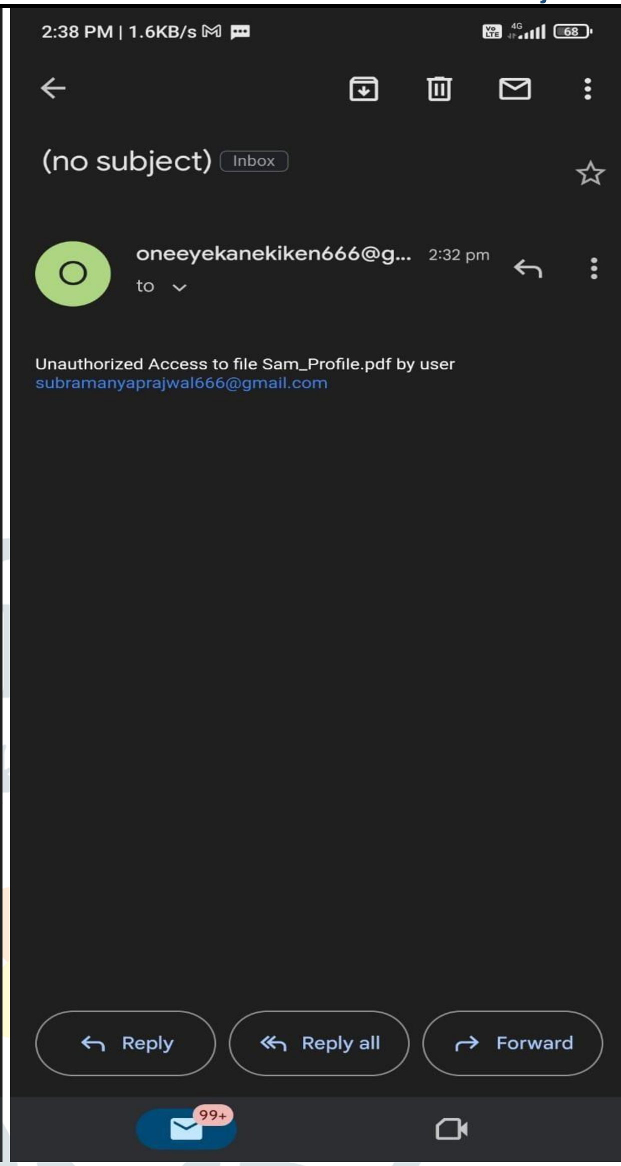

Figure 6.7: Unauthorized Access Notification

SYSTEM TESTING

Introduction

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionalityof components, sub-assemblies, assemblies and/or a finished product It is the process of exercisingsoftware with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of tests. Each test type addresses a specific testing requirement.

Types of tests

- **Unit testing**

Unit testing involves the design of test cases that validate that the internal program logic isfunctioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately tothe documented specifications and contains clearly defined inputs and expected results.

- **Integration testing**

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individuallysatisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

- **Functional testing**

Functional tests provide systematic demonstrations that functions tested are available asspecified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items

Valid Input : identified classes of valid input must be accepted. Invalid Input : identified classes of invalid input must be rejected.Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions,or special test cases. In addition, systematic coverage pertaining to identify Business process flows;data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current testsis determined.

System Testing

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is theconfiguration-oriented system integration test. System testing is based on process descriptions andflows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of theinner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, mustbe written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box. you cannot "see" into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as twodistinct phases. Unit texting also known as component testing refers to tests that verify the functionality of a specific code usually at the functional level.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Table 7.1: Test Cases

Test CaseID	Test Case Description	Preconditions	Steps	Expected Outcome	Status
TC01	User registration with validdetails	User is on the registration page	Enter valid details and submit	User account is created	Pass
TC02	User registration with existing username	User is on the registration page	Enter existing username and submit	Error message isdisplayed	Pass
TC03	User login with correct credentials	User is on the login page	Enter correct username and password and submit	Successful login	Pass
TC04	User login with incorrect credentials	User is on the login page	Enter incorrect username or password and submit	Error message isdisplayed	Pass
TC05	Admin login with correct credentials	Admin is on the login page	Enter correct admin username and password and submit	Admin panelis accessed	Pass
TC06	Admin login with incorrect credentials	Admin is on the login page	Enter incorrect admin username or password and submit	Error message isdisplayed	Pass
TC07	File upload by user	User is loggedin and on the file upload page	Upload a file	File is uploaded successfully	Pass
TC08	File download by authorized user	User is loggedin and file access granted	Request to download a file	File is downloaded successfully	Pass
TC09	File download by unauthorized user	User is loggedin but file access denied	Request to download afile	Access denied message displayed	Pass
TC10	Applying watermark to file	User is loggedin and on the file management page	Apply watermark to a file	Watermark is added successfully	Pass
TC11	User sets access permission for a file	User is loggedin and on the access control page	Configure access permissionsfor a file	Permissions updated successfully	Pass

TC12	Admin modifies useraccount	Admin is logged in and on the user management page	Edit userdetails	User details updated successfully	Pass
TC13	Real-time monitoring of user activity	Admin is logged in and on the monitoring page	Observe user activity	Activity log updated successfully	Pass
TC14	Key validation for authorized action	User has a valid accesskey	Attempt an authorized action	Action performed successfully	Pass
TC15	Key validation for unauthorized action	User has an invalid or expired access key	Attempt an unauthorized action	Access denied message displayed	Pass
TC16	Generating user activityreport	Admin is logged in and on the reporting page	Request a user activity report	Report generated successfully	Pass
TC17	System scalability test	System is under increased load	Simulate concurrent user interactions	System maintains performance	Pass
TC18	Encryption and decryption test	Data is prepared forencryption	Encrypt andthen decryptdata	Data remains secure	Pass
TC19	Compatibility test across browsers	User accesses the system using different browsers	Interact with the system using different browsers	Consistent performance observed	Pass
TC20	Error handling andedge cases	System is in different states	Trigger various error scenarios and edge cases	Appropriate error messages displayed	Pass

CONCLUSION AND FUTURE ENHANCEMENT

Conclusion

In conclusion, the development of the Data Leakage Detection System represents a significant stride in addressing the critical challenges posed by data breaches and unauthorized information access. Through a meticulous process of requirement analysis, system design, implementation, andtesting, we have strived to create a robust and user-centric solution that empowers organizations to safeguard their sensitive data in an increasingly interconnected world. This project's journey has underscored the imperative of proactive data security measures, considering the potential financial, reputational, and legal repercussions of data leaks. By offering a comprehensive systemwith features such as access control, real-time monitoring, and key validation, we aim to provide a tool that not only detects potential breaches promptly but also prevents unauthorized access andleakage in the first place. The collaboration between developers, users, and administrators has been instrumental in shaping the system to meet the diverse requirements and challenges. The integration of encryption, hashing, and watermarking techniques further enhances the security layers, fostering a trustworthy digital environment for data management and sharing. As the technological landscape continues to evolve, the Data Leakage Detection System's contribution todata protection and leak prevention stands as a testament to our commitment to a secure digital future. Through continued updates, user feedback, and ongoing improvements, we aspire to upholdthe system's efficacy in addressing emerging threats and maintaining data integrity.

Future Enhancement

- Advanced Machine Learning Algorithms: Integrating machine learning algorithms can enhance the system's ability to detect anomalies and predict potential data leakage incidentsbased on historical patterns.
- Behavioral Analytics: Incorporating user behavior analytics can help identify unusual patterns in user interactions, aiding in the early detection of insider threats or compromisedaccounts.
- Integration with Cloud Services: Future versions could include integration with popular cloud storage platforms, allowing users to monitor and secure files stored on external cloudservices.
- Automated Response Mechanisms: Implementing automated responses to potential threats,such as suspending user accounts or restricting file access, can enhance the system's abilityto react swiftly to security incidents.
- Mobile Application: Developing a mobile app could enable users and administrators to monitor and manage system activities on-the-go, enhancing accessibility and user experience.

REFERENCES

[1]S. Czerwinski, R. Fromm, and T. Hodes, “Digital Music Distribution and Audio Watermarking,”http://www.scientificcommons.org/430256 58, 2007. Available at: www.researchpublications.org NCAICN-2013, PRMITR,Badnera 399

[2]Y. Li, V. Swarup, and S. Jajodia, “Fingerprinting Relational Databases: Schemes and Specialties,” IEEE Trans. Dependable and Secure Computing, vol. 2, no. 1, pp. 34-45, Jan.-Mar. 2015.

[3]Y. Cui and J. Widom, “Lineage Tracing for General Data Warehouse Transformations,” The VLDB J., vol. 12, pp. 41- 58, 2014.

[4]Panagiotis Papadimitriou and Hector Garcia-Molina, “Data Leakage Detection,” IEEE Trans,Knowledge and Data Engineering, vol. 23, no. 1, January 2013.

[5]P. Bonatti, S.D.C. di Vimercati, and P. Samarati, “An Algebra for Composing Access ControlPolicies,” ACM Trans. Information and System Security, vol. 5, no. 1, pp.1- 35, 2011.

[6]Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System.

Link: https://bitcoin.org/bitcoin.pdf

[7]Buterin, V. (2014). Ethereum: A Next-Generation Smart Contract and Decentralized Application Platform.

Link: https://ethereum.org/whitepaper/

[8]Wood, G. (2014). Ethereum: A Secure Decentralized Generalized Transaction Ledger.

Link: https://ethereum.github.io/yellowpaper/paper.pdf

[9]Mougayar, W. (2016). The Business Blockchain: Promise, Practice, and Application of the Next Internet Technology.

[10] Tapscott, D., & Tapscott, A. (2016). Blockchain Revolution: How the Technology Behind Bitcoin is Changing Money, Business, and the World.

[11] Gervais, A., Karame, G. O., Wüst, K., & Capkun, S. (2016). On the Security and Performanceof Proof of Work Blockchains.

Link: https://eprint.iacr.org/2016/555.pdf

[12] Bonneau, J., Miller, A., Clark, J., Narayanan, A., Kroll, J. A., & Felten, E. W. (2015). SoK:Research Perspectives and Challenges for Bitcoin and Cryptocurrencies.

Link: https://www.cs.princeton.edu/~arvindn/publications/mining_CCS.pdf

[13] Zohar, A. (2015). Bitcoin: under the Hood.

Link: https://arxiv.org/abs/1507.06530.pdf

[14] Eyal, I., & Sirer, E. G. (2014). Majority is not enough: Bitcoin mining is vulnerable.

Link: https://arxiv.org/abs/1311.0243.pdf

[15] Kosba, A., Miller, A., Shi, E., Wen, Z., & Papamanthou, C. (2016). Hawk: The Blockchain Model of Cryptography and Privacy-Preserving Smart Contracts.

Link: https://eprint.iacr.org/2015/675.pdf