International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 09 | Sep 2022 www.irjet.net p-ISSN: 2395-0072

A System for Detection and Prevention of Data Leak

Aishwarya Jadhav[1], Prof. Pramila M. Chawan[2]

[1]M.Tech Student, Department of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India
[2]Associate Professor, Department of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India

---

Abstract - *Technology is growing exponentially in the recent years and most of the organizations store their data in digital format. With the rapid growth in technology, there is a need for maintaining security of 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 security purposes, 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 the information 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.*

Key Words: **Data Leak Prevention, Sensitive Data, Data Leak**

1. INTRODUCTION

Security has become an important factor in our life. Security is required in all sectors of industry. An attacker has various methods to access the confidential information of any organization. Hence, preventing such attackers from accessing the information is the main aim of information security. We need to implement various strategies to secure the information.

Data leak occurs when unauthorized users can access the confidential or sensitive data to. Data leak can happen intentionally through employees of the organization or malicious attackers. It can also be an unintentional leak by employees. In any case, the data is transferred outside the organization. Data leak usually occurs through email. It can also occur through data storage devices such as laptops.

Data is one of the most precious asset. Therefore, the prevention of data leak is the most important task for any organization.

Even with security measures like firewalls, data leak still occurs. In any organization, the employees have access to sensitive data. Hence, there is a chance that the data leak occurs through employees rather than through malicious attackers.

1.1 Types of Data

Any organization must deal with three types of data to prevent data leak:

1. **Data in motion**

It refers to data that is moving from the network to the outside world through the internet.

2. **Data at rest**

It refers to data that is stored in the file systems, databases and other storage methods

3. **Data at the endpoint**

It refers to data present at the endpoints in the network

Most of the organizations scan the emails that have been received from outside the organization for any malicious malwares. But, they do not check the emails sent outside the organization, thereby allowing the sensitive information to be sent outside the organization.

Most common causes of Data Breach are:

1. Hacking
2. Malware
3. Unintended Disclosure
4. Virus
5. Worms
6. Insider leak
7. Data loss

1.2 Causes of Data Leak

1. Virus Attack

If a machine is infected by viruses and worms, spyware, adware, etc. this might result in corruption and loss of data. In order to prevent this, anti-virus should be used.

2. Malicious Attack

Ill-intentioned and malicious attackers can hack into the system and steal, modify or delete valuable information. This will cause data leak.

Data leak prevention (DLP) is the practice for detection and prevention of data breaches and destruction of sensitive data. It is a set of tools and processes used to ensure that unauthorized users do not access, delete or modify the sensitive data.

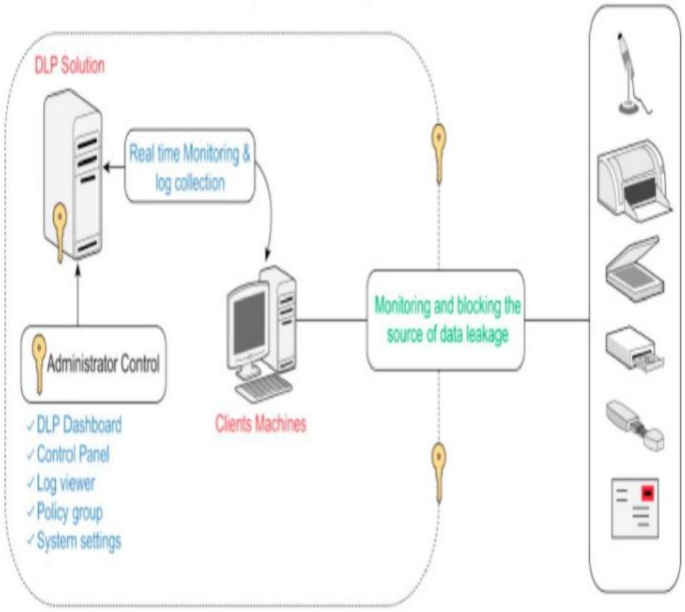


Fig -1: Data Leak Prevention model

DLP software classifies data into different categories and checks for violations of policies defined by organizations.

Once these policies are violated, DLP issues alerts, and other protective measures to prevent end users from accidentally or maliciously sharing confidential information.

1.3 Types of Data Leak Prevention System

The DLP can be classified into two types:

- Network DLP
- Endpoint DLP

➢ **Network DLP**

Network DLP stands for Network Data Loss Prevention. It is a technique for protecting communications over the network such as web applications, emails, and data transfer mechanisms of the organization.

It helps to prevent the loss of sensitive data on the network.
Moreover, it allows the company to encrypt data and to block risky information flows in order to monitor and control the flow of data over the network according to the regulatory compliance.

➢ **Endpoint DLP**

Endpoint DLP stands for Endpoint Data Leak Prevention. It protects sensitive data at the endpoints. It also helps the organization to track employee behaviors.

It monitors and addresses daily risky actions like sending emails, uploading data to the cloud, etc. It provides a wider range of threat protection.

In this project, we will be implementing Endpoint DLP

2. PROPOSED METHODOLOGY

2.1 Problem Statement

To develop a system for Detection and Prevention of Data Leak.

2.2 Problem Elaboration

In this system, the data will be secured by using DLP. DLP is a method to prevent the users from sending sensitive data outside the organization. The system is used to check for any activities of data transfer that might lead to data leak.

The system will control and monitor endpoint activities. It will also monitor the data in the cloud to protect data at rest, in motion, and in use. It will also generate reports and identify weaknesses in the system to enhance the security.

2.3. Proposed System

The proposed system suggests a System for Detection and Prevention of Data Leak. The system will constantly monitor the activities of the employees and will restrict any malicious activities. If any suspicious activity from employees is found, the system will inform the admin through generation of incident mail.

2.3.1 Parameters to Detect and Prevent Data Leak:

The system uses a set of parameters to detect and prevent Data Leak. They are as follows:

- Time Restriction
- Extension Restriction
- Keywords Restriction

1) Time Restriction

It is a technique which will restrict any malicious activity from users by applying time constraints. Here, the admin of the system can decide a particular time frame for sending the mails. The mail can be sent only in the time frame defined by the admin. It will not be sent outside the time frame.

e.g.: The time restriction can be set from 8:00 am to 9:00 pm. Any transmissions outside this time frame will trigger incident mail to the admin.

2) Extension Restriction

It is a technique which will restrict the files with extensions that cannot be read by the system. The following are the readable file formats for the system:

- Excel (.xls)
- Word (.doc)
- PDF (.pdf)
- Text File (.txt)

Transmission of files with extensions other than Excel (.xls), Word (.doc), PDF (.pdf) and Text File (.txt) will pose a threat to the confidentiality of the information of any organization.

Hence, the Extension Restriction feature allows the admin to block the transmission of files with all other extensions like Jpg, jpeg, png, mp3, mp4, etc. which cannot be read by the system.

3) Keywords Restriction

It is a technique which will restrict the transmission of files which includes any confidential data.

The admin can define a set of suspicious keywords. The system will check whether the file contains any of the keywords.

If the file contains the suspicious keywords, then the system will block the transmission, else it will allow the transmission.

2.3.2 System Architecture

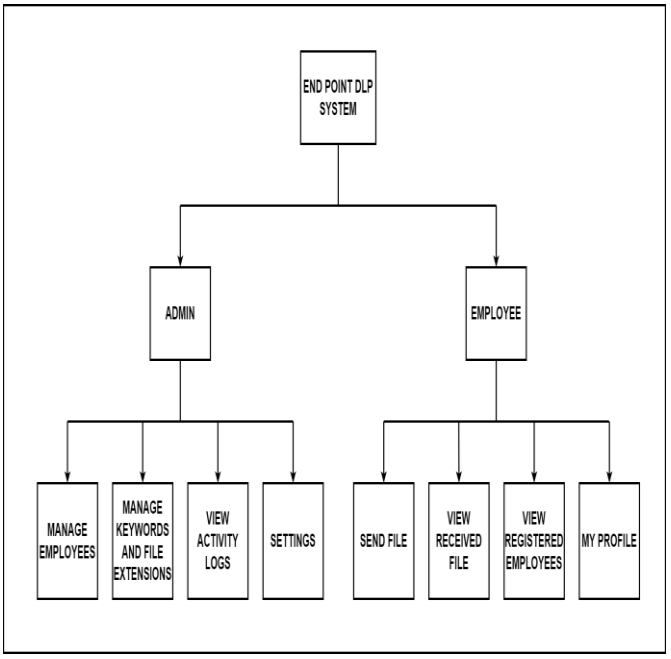


Figure 2. System Architecture

The system has two major modules:

- ADMIN MODULE
- EMPLOYEE MODULE

- **ADMIN MODULE:**

The Admin can login using the credentials provided by the system.

The functionalities of Admin Module are:

1. Manage Employees:

The admin has the rights to add, update and delete employee details. After registration of the user, login credentials will be generated and will be sent via mail.

2. Manage Keywords & File Extensions:

The admin can define a set of keywords. The keywords can be added, modified and deleted. Admin will block the non - readable file Extensions.

3. View Activity Logs:

The activity logs will be displayed here.

4. Settings:

The admin can set a time frame for sending emails. Admin will also provide Email id to receive incident email.

➢ **EMPLOYEE MODULE:**

The employee can login using the credentials provided by the system.

The functionalities of Employee Module are:

1. Send File:

The employee can send files to anyone inside or outside the organization.

2. View Received Files:

The employee can download the files received through email.

3. View Registered Employees:

The employee can view the profile of other employees and get information like their email id.

4. My Profile:

The employees can view and edit their own profile.

2.3.3 Working

The system constantly monitors the activities of employees to check whether there is any malicious activity or not. The employee can send files to anyone inside or outside the organization.

Whenever an employee wants to send a file over mail, the system will check all the parameters and will either block the mail or allow the mail to be sent.

The system will identify data leak based on the parameters used to identify Data Leak, i.e. Time Restriction, Extension Restriction and Keywords Restriction.

The system will also generate incident mail to the admin and create an activity log.

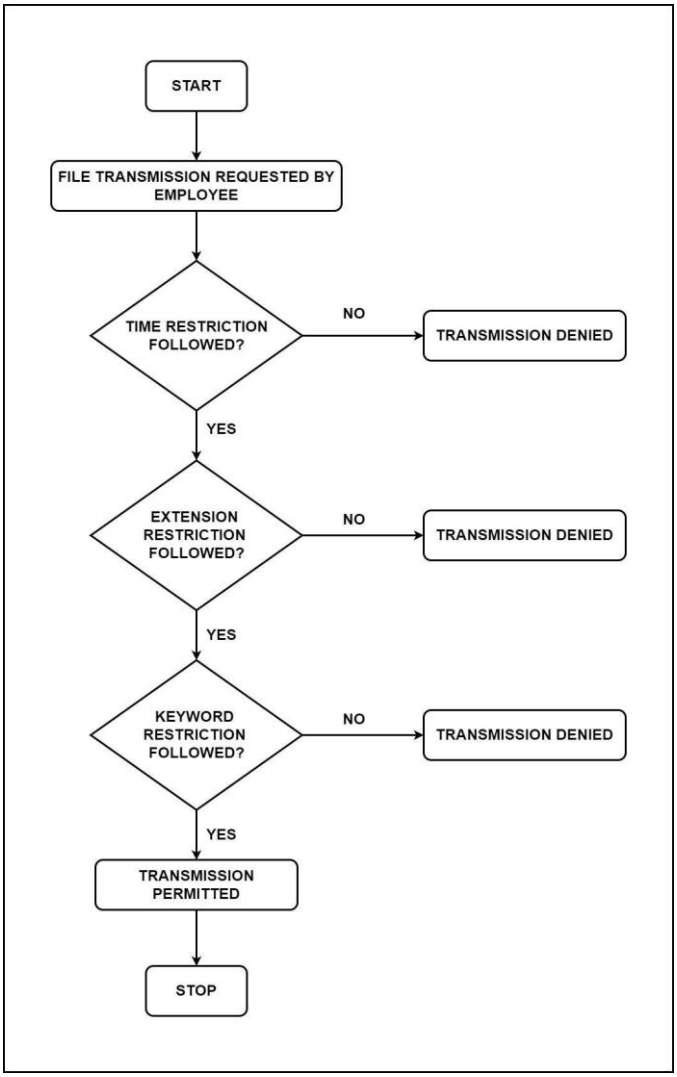


Figure 3. Working

3. RESULTS

The following are the various case scenarios in which the system is able to restrict the transmission of suspicious data:

CASE 1: Restricted due to Time:

In this case, the transmission outside the timeframe mentioned in the system is blocked.

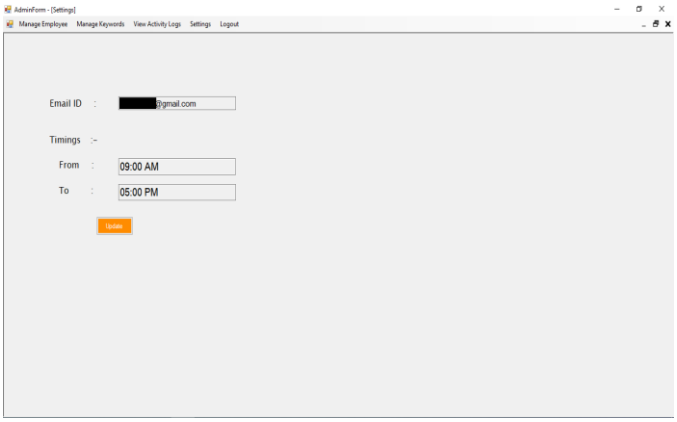


Figure 4: Settings

As we can see in the figure, the time set by us is from 9:00 am to 5:00 pm.

Now when we try to send a file outside the time frame, the system blocks the transmission.

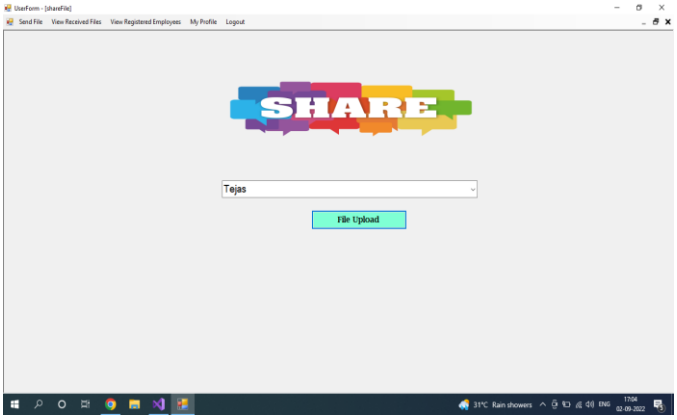


Figure 5: Transmission Blocked

An incident email is generated to the admin and the activity logs are recorded.

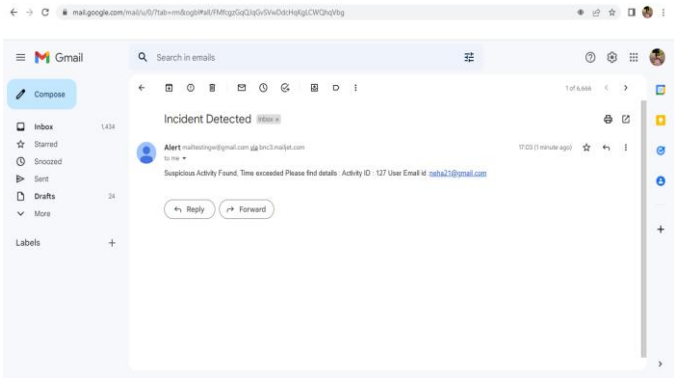


Figure 6: Incident Mail

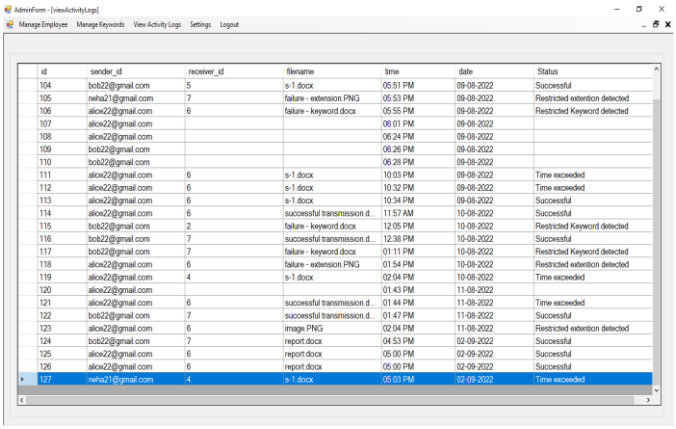


Figure 7: Activity Logs

CASE 2: Restricted due to Extension:

In this case, the transmission of the file is carried out within the timeframe mentioned in the system, but the extension for the file is not readable.

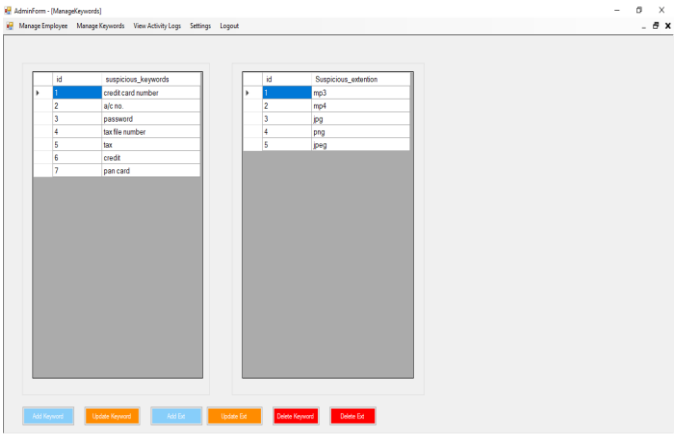


Figure 8: Manage Files and Extensions

As we can see in the figure, the approved file extensions for transmission are:

- Excel (.xls)
- Word (.doc)
- PDF (.pdf)
- Text File (.txt)

Now when we try to send a file within the time frame but with a suspicious extension, the system blocks the transmission.

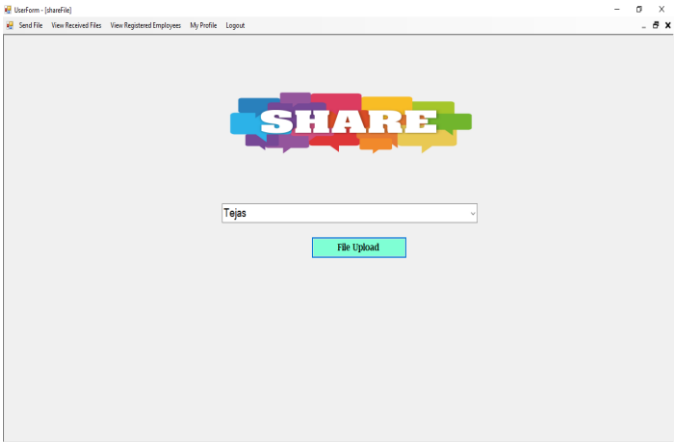


Figure 9: Transmission Blocked

An incident email is generated to the admin and the activity logs are recorded.

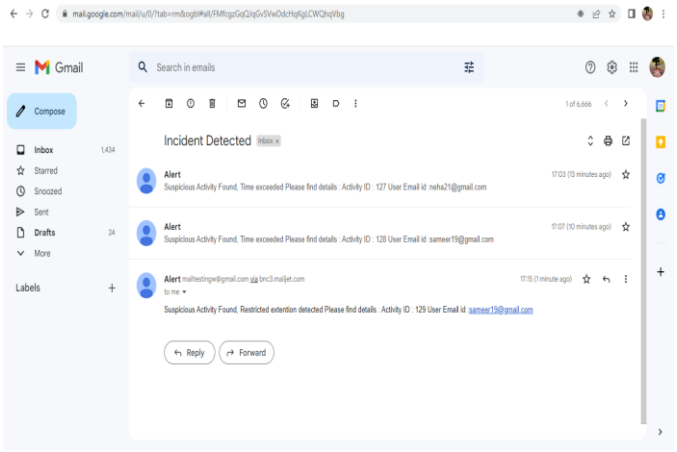
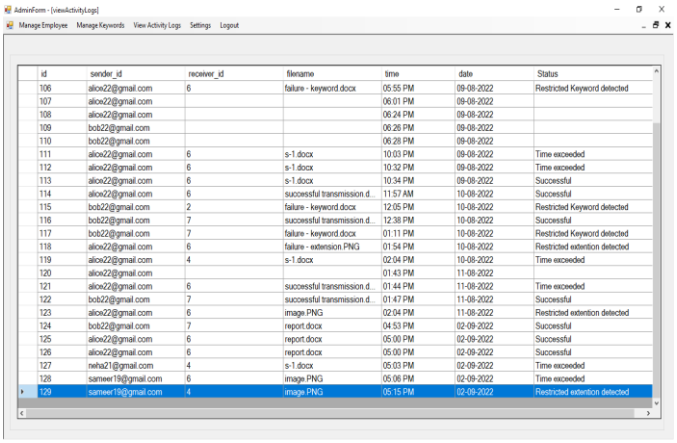


Figure 10: Incident Mail



id	sender_id	receiver_id	filename	time	date	Status
106	alice22@gmail.com	6	failure - keyword.docx	05:55 PM	09-08-2022	Restricted Keyword detected
107	alice22@gmail.com			06:01 PM	09-08-2022	
108	alice22@gmail.com			06:24 PM	09-08-2022	
109	bob22@gmail.com			06:26 PM	09-08-2022	
110	bob22@gmail.com			06:28 PM	09-08-2022	
111	alice22@gmail.com	6	s-1.docx	10:03 PM	09-08-2022	Time exceeded
112	alice22@gmail.com	6	s-1.docx	10:32 PM	09-08-2022	Time exceeded
113	alice22@gmail.com	6	s-1.docx	10:34 PM	09-08-2022	Successful
114	alice22@gmail.com	6	successful transmission.d...	11:57 AM	10-08-2022	Successful
115	bob22@gmail.com	2	failure - keyword.docx	12:05 PM	10-08-2022	Restricted Keyword detected
116	bob22@gmail.com	7	successful transmission.d...	12:38 PM	10-08-2022	Successful
117	bob22@gmail.com	7	failure - keyword.docx	01:11 PM	10-08-2022	Restricted Keyword detected
118	alice22@gmail.com	6	failure - extension.PNG	01:54 PM	10-08-2022	Restricted extention detected
119	alice22@gmail.com	4	s-1.docx	02:04 PM	10-08-2022	Time exceeded
120	alice22@gmail.com			01:43 PM	11-08-2022	
121	alice22@gmail.com	6	successful transmission.d...	01:44 PM	11-08-2022	Time exceeded
122	bob22@gmail.com	7	successful transmission.d...	01:47 PM	11-08-2022	Successful
123	alice22@gmail.com	6	image.PNG	02:04 PM	11-08-2022	Restricted extention detected
124	bob22@gmail.com	7	report.docx	04:53 PM	02-09-2022	Successful
125	alice22@gmail.com	6	report.docx	05:00 PM	02-09-2022	Successful
126	alice22@gmail.com	6	report.docx	05:00 PM	02-09-2022	Successful
127	neha21@gmail.com	4	s-1.docx	05:03 PM	02-09-2022	Time exceeded
128	sameer19@gmail.com	6	image.PNG	05:06 PM	02-09-2022	Time exceeded
129	sameer19@gmail.com	4	image.PNG	05:15 PM	02-09-2022	Restricted extention detected

Figure 11: Activity Logs

CASE 3: Restricted due to Keyword:

In this case, the transmission of the file is carried out within the timeframe mentioned in the system and the extension for the file is readable, but there are keywords present in the file.

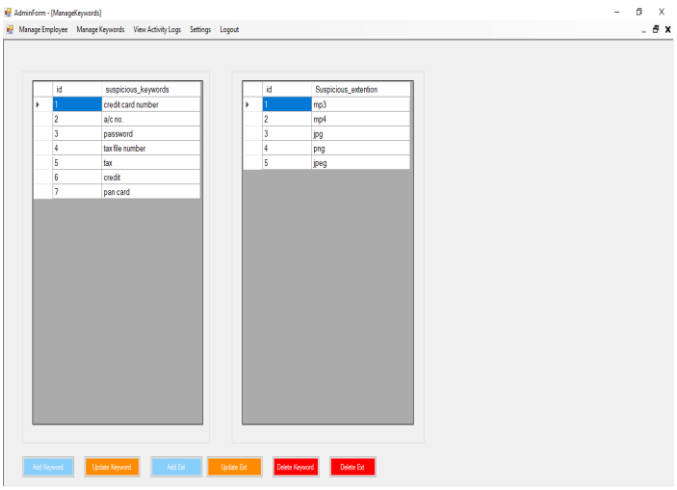


Figure 12: Manage Files And Extensions

As we can see in the figure, the given set of keywords are restricted.

Now when we try to send a word file with suspicious keywords, the system blocks the transmission.

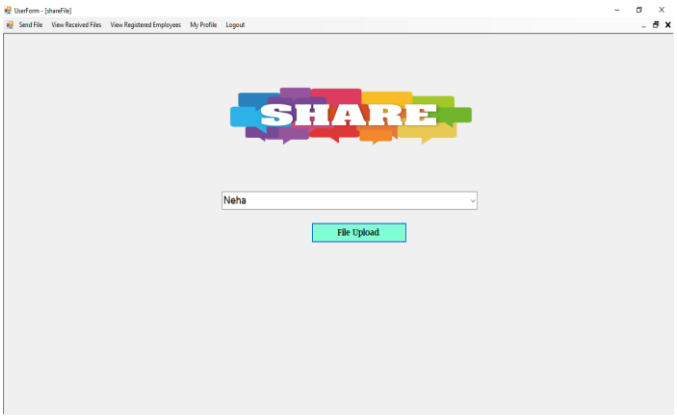


Figure 13: Transmission Blocked

An incident email is generated to admin and the activity logs are recorded.

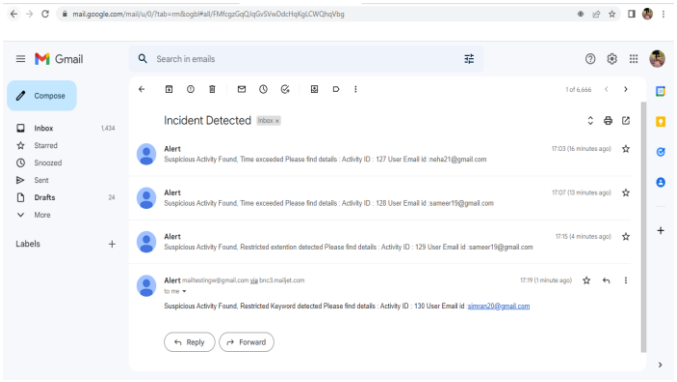
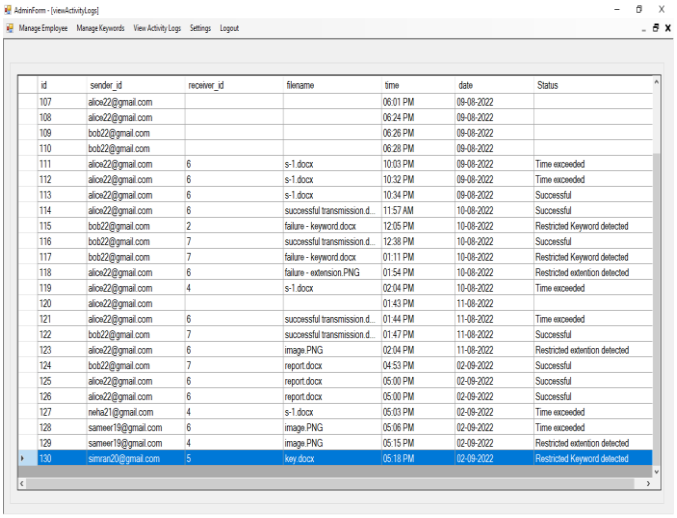


Figure 14: Incident Mail



id	sender_id	receiver_id	filename	time	date	Status
107	alice22@gmail.com			06:01 PM	09-08-2022	
108	alice22@gmail.com			06:24 PM	09-08-2022	
109	bob22@gmail.com			06:26 PM	09-08-2022	
110	bob22@gmail.com			06:28 PM	09-08-2022	
111	alice22@gmail.com	6	s-1.docx	10:03 PM	09-08-2022	Time exceeded
112	alice22@gmail.com	6	s-1.docx	10:32 PM	09-08-2022	Time exceeded
113	alice22@gmail.com	6	s-1.docx	10:34 PM	09-08-2022	Successful
114	alice22@gmail.com	6	successful transmission.d...	11:57 AM	10-08-2022	Successful
115	bob22@gmail.com	2	failure - keyword.docx	12:05 PM	10-08-2022	Restricted Keyword detected
116	bob22@gmail.com	7	successful transmission.d...	12:38 PM	10-08-2022	Successful
117	bob22@gmail.com	7	failure - keyword.docx	01:11 PM	10-08-2022	Restricted Keyword detected
118	alice22@gmail.com	6	failure - extension.PNG	01:54 PM	10-08-2022	Restricted extention detected
119	alice22@gmail.com	4	s-1.docx	02:04 PM	10-08-2022	Time exceeded
120	alice22@gmail.com			01:43 PM	11-08-2022	
121	alice22@gmail.com	6	successful transmission.d...	01:44 PM	11-08-2022	Time exceeded
122	bob22@gmail.com	7	successful transmission.d...	01:47 PM	11-08-2022	Successful
123	alice22@gmail.com	6	image.PNG	02:04 PM	11-08-2022	Restricted extention detected
124	bob22@gmail.com	7	report.docx	04:53 PM	02-09-2022	Successful
125	alice22@gmail.com	6	report.docx	05:00 PM	02-09-2022	Successful
126	alice22@gmail.com	6	report.docx	05:00 PM	02-09-2022	Successful
127	neha21@gmail.com	4	s-1.docx	05:03 PM	02-09-2022	Time exceeded
128	sameer19@gmail.com	6	image.PNG	05:06 PM	02-09-2022	Time exceeded
129	sameer19@gmail.com	4	image.PNG	05:15 PM	02-09-2022	Restricted extention detected
130	simran20@gmail.com	5	key.docx	05:18 PM	02-09-2022	Restricted Keyword detected

Figure 15: Activity Logs

CASE 4: Successful Transmission:

In this case, the transmission of the file is carried out within the timeframe mentioned in the system, the file has appropriate extension and it does not contain any keywords.

Here, the system allows the transmission and the activity logs are recorded.

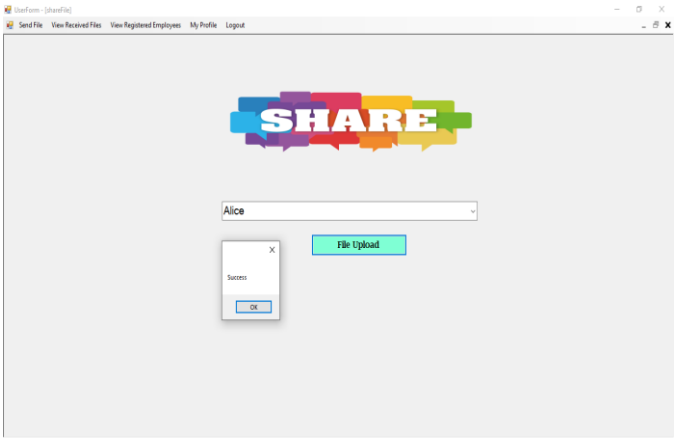
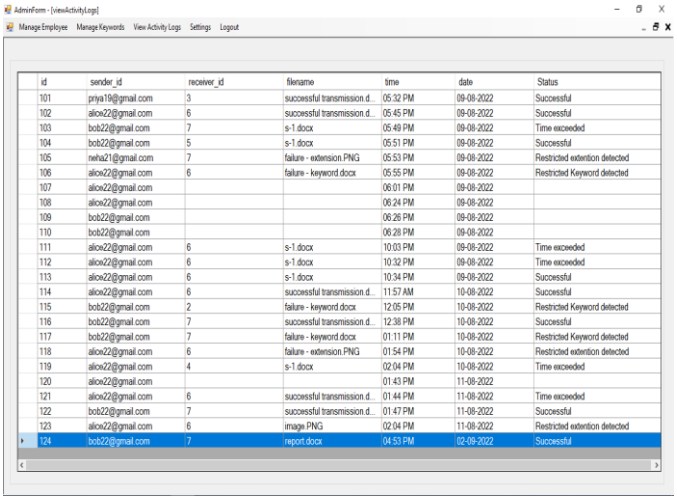


Figure 16: Transmission Successful

As we can see,the system shows a success message.



id	sender_id	receiver_id	filename	time	date	Status
101	priya19@gmail.com	3	successful transmission.d...	05:32 PM	09-08-2022	Successful
102	alice22@gmail.com	6	successful transmission.d...	05:45 PM	09-08-2022	Successful
103	bob22@gmail.com	7	s-1.docx	05:49 PM	09-08-2022	Time exceeded
104	bob22@gmail.com	5	s-1.docx	05:51 PM	09-08-2022	Successful
105	neha21@gmail.com	7	failure - extension.PNG	05:53 PM	09-08-2022	Restricted extention detected
106	alice22@gmail.com	6	failure - keyword.docx	05:55 PM	09-08-2022	Restricted Keyword detected
107	alice22@gmail.com			06:01 PM	09-08-2022	
108	alice22@gmail.com			06:24 PM	09-08-2022	
109	bob22@gmail.com			06:26 PM	09-08-2022	
110	bob22@gmail.com			06:28 PM	09-08-2022	
111	alice22@gmail.com	6	s-1.docx	10:03 PM	09-08-2022	Time exceeded
112	alice22@gmail.com	6	s-1.docx	10:32 PM	09-08-2022	Time exceeded
113	alice22@gmail.com	6	s-1.docx	10:34 PM	09-08-2022	Successful
114	alice22@gmail.com	6	successful transmission.d...	11:57 AM	10-08-2022	Successful
115	bob22@gmail.com	2	failure - keyword.docx	12:05 PM	10-08-2022	Restricted Keyword detected
116	bob22@gmail.com	7	successful transmission.d...	12:38 PM	10-08-2022	Successful
117	bob22@gmail.com	7	failure - keyword.docx	01:11 PM	10-08-2022	Restricted Keyword detected
118	alice22@gmail.com	6	failure - extension.PNG	01:54 PM	10-08-2022	Restricted extention detected
119	alice22@gmail.com	4	s-1.docx	02:04 PM	10-08-2022	Time exceeded
120	alice22@gmail.com			01:43 PM	11-08-2022	
121	alice22@gmail.com	6	successful transmission.d...	01:44 PM	11-08-2022	Time exceeded
122	bob22@gmail.com	7	successful transmission.d...	01:47 PM	11-08-2022	Successful
123	alice22@gmail.com	6	image.PNG	02:04 PM	11-08-2022	Restricted extention detected
124	bob22@gmail.com	7	report.docx	04:53 PM	02-09-2022	Successful

Figure 17: Activity Logs

The receiver can download and view the file.

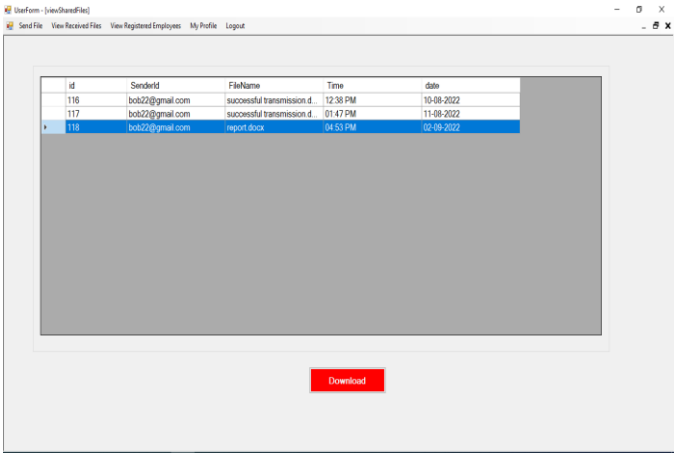


Figure 18: Download Recieved Files

4. CONCLUSION

Data leak is a major issue for many organizations. Data leak can have a disastrous effect on any organization. Hence, preventing data leak is very important. In this paper, a system is proposed for detection and prevention of data leak, which will achieve the security goals of an organization. The proposed method is easy to implement and can be useful for many organizations.

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 Control Policies," ACM Trans. Information and System Security, vol. 5, no. 1, pp.1-35, 2011.

BIOGRAPHIES

Aishwarya Jadhav is currently pursuing M. Tech from VJTI COE, Mumbai. She has done her B.E. (Computer Engineering) from Atharva College of Engineering.

Prof. Pramila M. Chawan, is working as an Associate Professor in the Computer Engineering Department of VJTI, Mumbai. She has done her B.E.(Computer Engineering) and M.E.(Computer Engineering) from VJTI College of Engineering, Mumbai University.
She has 28 years of teaching experience and has guided 85+ M. Tech. projects and 130+ B. Tech. projects. She has published 143 papers in the International Journals, 20 papers in the National/International Conferences/ Symposiums. She has worked as an Organizing Committee member for 25 International Conferences and 5 AICTE/MHRD sponsored Workshops/STTPs/FDPs. She has participated in 16 National/International Conferences. Worked as Consulting Editor on –
JEECER, JETR, JETMS, Technology Today, JAM&AER Engg. Today, The Tech. World
Editor – Journals of ADR
Reviewer -IJEF, Inderscience
She has worked as NBA Coordinator of the Computer Engineering Department of VJTI for 5 years. She had written a proposal under TEQIP-I in June 2004 for 'Creating Central Computing Facility at VJTI'. Rs. Eight Crore were sanctioned by the World Bank under TEQIP-I on this proposal. Central Computing Facility was set up at VJTI through this fund which has played a key role in improving the teaching learning process at VJTI.
Awarded by SIESRP with Innovative & Dedicated Educationalist Award Specialization : Computer Engineering & I.T. in 2020
AD Scientific Index Ranking (World Scientist and University Ranking 2022) –
2nd Rank- Best Scientist, VJTI Computer Science domain
1138th Rank- Best Scientist, Computer Science, India