Master of Science in Internetworking

Department of Computing Science

MINT 709 Capstone Project Report

Analysis of Best Practices for Data Leak Management and Prevention of Data Harvesting

Under the Guidance of

Leonard Rogers

Submitted By

Abhijith Daggupati

Fall 2020

Acknowledgement

I would like to thank and express my gratitude to my mentor, *Mr. Leonard Rogers*, for his extraordinary guidance and suggestions throughout this project, which helped me complete this project.

I would like to thank my program coordinator, *Mr. Shahnawaz Mir*, for supporting me throughout my program.

I would like to thank my project supervisor, *Dr. Mike MacGregor*, for his support and guidance in this project.

Finally, I would like to thank my family and friends who supported me throughout this entire project.

Table of Content

List of Figures

List of Tables

Abstract

In this technological world, data leakage poses a severe threat to every organization as well as to individuals. Every day, a new data leak or a data breach is happening, and the number of data leaks is increasing continuously, resulting in the rise in the cost of handling a data leak. Whether caused due to an outside attacker or unintentional mistakes by the employees, a data leak will cause a destructive effect on the organization.

The data leak management techniques play an essential role in preventing and handling data leaks. Although these measures give a certain level of protection, the threat vectors are evolving continuously to bypass these security measures.

Nowadays, the number of websites worldwide has been growing continuously, containing a high volume of valuable data. Data harvesting became a big concern for the website owners. Moreover, it also affects the company's business in many ways. Despite data harvesting prevention technique exists, it is not easy to prevent the bots from scraping entirely.

The first part of this report aims at analysing various causes of data leaks and how they impact the company's business. And this paper provides the best practices for preventing data leaks based on each threat vector and the data leak handling techniques to mitigate the data leak to reduce the impact on the organization. The second part of this report focuses on various causes of data harvesting and the process of scraping data from the websites. Finally, it discusses the required actions to prevent data harvesting from happening again in the future.

Part I

Chapter 1: Introduction of Data Leaks

1.1 Data Leak

Data such as trade secrets, customer data and trading algorithms are critical assets of any organization, so the leakage of sensitive data causes a significant impact on that organization, both financially and reputationally. Thus, handling data carefully to avoid it from leakage has becoming a priority.

Many organizations use data leak management techniques to detect and prevent data from leaking and potential data destruction. Even though companies have been implementing data leakage prevention measures, it is still happening.

A Data leak is the accidental exposure of sensitive data from an organization to unauthorized people. The data leaks could happen in any organization due to insider or outsider threats. These threats will be discussed later in chapter 3.

While the terms data leak and data breach are often considered as similar, they are two different types of data exposures: [1]

- If sensitive data is exposed due to a successful attack, it is said to be a Data Breach.
- If sensitive data is exposed due to insufficient data security practices or accidental action by an individual, it is said to be a Data Leak. It does not require any Cyber-attack.

1.2 History and Evolution of Data Leaks

Data Leaks have been occurring continuously for the past few decades, but they also cause significant losses from time to time.

Data leaks and breaches did not start when organizations began to store their data in the digital form. These have existed from when the organizations maintained their data in paper records. Before computing became commonplace, a data leak could be as simple as making a file of information visible to anyone without authorization. [2]

While data leaks and breaches were occurring before 2005, most of the major leaks recorded in history are reported in 2005 or beyond. This has led to the exponential growth in the volume of data leaks year after year, giving cybercriminals an excellent opportunity to exploit massive volumes of data. [2]

Let’s review some severe data leaks that have occurred since 2005, whose impact was very high.

1.2.1 Yahoo

2012-2014 – 500 million to 3 billion users

In September 2016, Yahoo announced that in 2014 it had been become a victim of one of the biggest data breaches in history. They mentioned that hackers compromised the names, email addresses, date of birth, telephone numbers and the encrypted passwords of 500 million users. [3]

Later in December 2016, they disclosed another breach from 2013 by a different attacker that compromised the names, dates of birth, email addresses, encrypted passwords, and security questions and answers of 1 billion user accounts. Yahoo, which was acquired by the Verizon communications, revised that estimate in October 2017 to include all of its 3 billion user accounts. Both these breaches occurred due to poor security practices followed by Yahoo. [3] [4]

In 2017, the officials stated that a Russian hacker Alexsey Belan used a spear phishing attack to gain access to the yahoo network. They explained that it came from a single user in Yahoo’s corporate office. The employee was sent a spear-phishing mail with a link, as soon as they clicked, it downloaded malware onto the network. Later, a backdoor was created on the Yahoo server to get more privileged access to the Yahoo email account’s internal control center. The attacker then copied and exported a backup of Yahoo’s User Database, which he used to gain all account holders details. [4]

Credentials were then forged to trick Yahoo servers into recognizing them as genuine account holders who had mostly stayed logged in. The manoeuvre, appetizingly called “cookie minting”, allowed them to read the contents of some 6,500 Yahoo accounts without needing a password or username. [4]

1.2.2 Facebook

June 2013 - 6 million users

Facebook has been failing to protect user data for a long time. A series of data leaks occurred starting in 2013.

In June 2013, Facebook found a bug that causes personal data exposure, such as phone numbers and email addresses of 6 million users to unauthorized viewers for over a year. This technical glitch is said to have started in 2012 and was not noticed until 2013. Later Facebook fixed this bug and reported the leak to regulators. While it was not a major violation of the year, it marked the beginning of Facebook's problems when it came to its data security practices. [5]

May 2018- 14 million users

On Facebook, we can use different privacy settings on our posts and profile to manage who can view our posts, specifically our friends list or the public. A technical glitch occurred in May 2018 caused the private posts of 14 million people to be shared openly without their consent. However, this bug was active for only five days; Facebook immediately restored all posts to their typical security settings. [5]

September 2018- 50 million users

In September 2018, Facebook announced that around 50 to 90 million user accounts were hacked and accessed by the attackers, allowing them to see everything on a user's profile. They also confirmed that users logged in to third-party sites using their Facebook account could also be affected.

The situation turned out to be complicated and relied on three different bugs on the platform related to the Facebook feature 'View As' that allows people to see what their profile looks like to someone else. [5]

The first bug in the system caused the Facebook video upload tool to appear on the "View As" page. The second bug then promoted the video uploader to create a login token, giving the attackers the same login permissions as the Facebook mobile application.

Finally, when the video uploader appeared in "View As" mode, the access code was provided to the attacker. In response, Facebook removed 90 million users across all platforms and asked them to log in and reset their passwords and temporarily deactivated the "View as" feature. [5]

March 2019- 600 million users

In March 2019, according to cybersecurity expert Brian Krebs report, 600 million user credentials were stored in the plain text files. Although only the employees have access to these files, around 2,000 Facebook employees can access them. They revealed that the passwords of

the millions of Instagram users had been stored in the plaintext. Facebook reiterated those passwords had not been leaked. [5]

April 2019- 1.5 million users

1.5 million new user's email contacts were harvested while creating Facebook accounts without the user's consent. The email contacts were imported automatically without permission after verifying the email address by entering the password. Moreover, users do not have the right to cancel this process. Then Facebook used this data to improve the performance of ads and in suggesting new friends. Even though they cannot see the email's content, they still determine whom you are communicating with, which is a big privacy concern. [5]

September 2019- 419 million users

Around 419 million Facebook user's data containing a unique Facebook ID, and contact numbers were found sitting on an exposed server which was not owned by Facebook. It is like the data leak which occurred in April 2019, in which 540 million data records had been found on a public server, which was hosted by a Mexican company, Cultura Colectiva. [5]

December 2019- 309 million users

Facebook ended 2019 with a bang while another database was left exposed. More than 300 million usernames, mobile numbers and User IDs were left unprotected on the dark web for nearly two weeks. According to the Security expert Bob Diachenko report, this breach resulted from an illegal scraping operation or Facebook API abuse by hackers in Vietnam. [5]

The number of those affected initially was 267 million. However, in March 2020, it was found that the second server contained 42 million records disclosed by the same criminal group, bringing the total to 309 million. Once again, it is unknown if anyone has been affected by the crime, but it certainly puts users at risk of spam and phishing attacks. [5]

Overall, in 2019 alone, more than 1 billion user data records were leaked, equal to half of Facebook's users.

1.2.3 Elasticsearch Server

October 2019- 1.2 billion users

On October 16, 2019, security researchers found an open Elasticsearch server that contained four terabytes of unprotected data containing 1.2 billion user records. These exposed data

records contain names, email addresses, contact numbers, Facebook and LinkedIn profile data. The data originated from two different data enrichment companies known as People Data Labs (PDL) and OxyData (OXY). The data found in the Elastic server is almost the same data matched with the data returned by the People Data Labs API except for the education histories. [6]

1.2.4 Equifax

September 2017- 147.9 million users

On September 7, 2017, one of the most significant credit bureaus in the US named Equifax, was leaked due to an application vulnerability on their website. The lack of proper system segmentation led to easy access and movement for the attackers and caused leakage of 147.9 million consumer's data. Including social security numbers, DOB, addresses and driving license numbers of 143 million consumers and credit card data of 209,000 consumers. [3]

1.2.5 Microsoft

January 2020- 250 million users

In January 2020, Comparitech security researchers led by Bob Diachenko discovered the leak and notified Microsoft that a server containing 250 million user records was exposed. The exposed data contained customer service and support logs, which included conversations between support agents and customers from 2005 to December 2019.

Although most of the Personally Identifiable Information (PII) was redacted from the records, some customer's information such as email addresses, IP addresses, support agent's emails, case numbers and resolutions were exposed, leading to the risk of Tech-support scams targeting Microsoft customers. Microsoft later declared that cybercriminals did not access the exposed information. Moreover, they have concluded this leak as a 'misconfiguration of an internal customer support database'. [7]

Chapter 2: Analysis of Data Leaks

2.1 Types of Data in Data Leaks

Data leaks are happening very often; since 2013, there are approximately 3,809,448 records stolen from breaches every day, 158,727 per hour, 2,645 per minute, and 44 every second of a daily report by Cybersecurity Ventures. However, there are several types of data being stolen or leaked due to cybercriminals.

2.1.1 Customer Information

Most cybercriminals mainly focus on the PII. This information is used to steal the customer's identity and can contain: [1]

- **Personally Identifiable Information:** name, mobile number, address, email address, username and password.
- **Activity Information:** order history, payment history and browsing data.
- **Protected Health Information:** Medical test reports, prescriptions and medical history.
- **Credit Card Information:** Credit card numbers, expiration dates, CVV and billing zip codes.

2.1.2 Company Information

Some of the data leaks include data related to corporate companies. Exposure of corporate data leads to a devastating impact on that company. Types of data involved can include: [1]

- **Internal Communications:** emails, memos and documents containing company operations details.
- **Metrics:** performance statistics, projections, and other related data.
- **Strategy:** Hierarchy, messaging details, company roadmaps, electronic rolodexes and other crucial business data.

2.1.3 Trade Secrets

Trade secrets are highly critical data for every organization. Without this data, every company cannot compete with its competitors. Trade secrets are the most dangerous thing to be disclosed in the data leaks. Exposure of this data results in considerable losses in the business. Trade secrets include: [1]

- **Formulas, Plans and Designs:** Secret Information related to present and upcoming products.
- **Software and Code:** Proprietary software and tools developed for commercial or personal usage.
- **Commercial Methods:** Market strategies and contacts.

2.1.4 Analytics

The massive data sets include multiple sources of information that reflect major picture trends, trajectories and patterns. This data helps to understand the individuals and to predict their next actions with accuracy. This may sound odd, but the analytical data could be used in the elections to sway voters, which was happened in the Cambridge Analytica Case study. So, data required to perform analytics is to be placed securely to mitigate risks. Analytics data can contain [1]

- **Psychographic Data:** Personality attributes, Preferences and Demographics.
- **Behavioural Data:** User behaviour details, during usage of websites and applications.
- **Modelled Data:** Predicted attributes obtained from gathered information.

2.2 Outcomes of Data Leak

It is essential to understand how cybercriminals exploit the data and use them for illegal activities. Five common ways for exploiting data are given below: [1]

- **Credit Card Fraud:** Cybercriminals use credit card information to commit frauds.
- **Selling Data on Black Markets:** When the data is exposed, it may be placed for auction on the dark web. Cybercriminals specialize in searching and finding unsecured cloud instances and vulnerable databases to obtain PII to sell those data to perform Identity Thefts, frauds and Phishing operations.
- **Extortion:** The data is held over for ransom on a company's head or to damage the company's reputation.
- **Degrading Competitive Advantage:** Competitors of the companies may take advantage of the data leaks. Sensible information of a company such as trade secrets and customer lists could make its competitors to know about their resources and strategies.

- **Use the Data Themselves:** Hackers rarely monetize the stolen data by using it themselves to make purchases or commit fraud. It is very rare, because it may attract the attention of the authorities.

2.3 Cost of Data Breach

It is essential to estimate the cost for a data breach in a company. Security institutes gather direct and indirect expenses due to data breaches to calculate the data breach's cost.

Direct expenses include forensic experts to study and investigate the breach and hotline support. Whereas indirect costs contain in-house investigations, communications and customer turnover. [8]

Accounting methods such as Activity-based costing estimate the cost by identifying the activities and assigning a cost based on actual use. Four process-related activities drive a range of expenditures associated with the data breach. They are: [8]

- **Detection and Escalation:** Activities conducted by the company to detect the data breach.
    - Investigation and forensic activities
    - Assessment services
    - Audit services
    - Crisis management
    - Communications to executives and boards
- **Notification:** Activities allowing the company to notify data protection regulators, data subjects and other third-party companies.
    - General notices, emails, letters, outbound calls to data subjects
    - Determination of regulatory requirements
    - Communication with regulators
    - Involvement of outside experts
- **Ex-post Response:** Activities to help victims of a data breach to communicate with the company and resolving activities to victims and regulators.
    - Help desk and inbound communication
    - Identity protection services and credit monitoring
    - Issuing new accounts
    - Legal expenditure

- Discount on products
- Regulatory fines

➢ **Lost Business:** Activities performed to minimize the loss of customers, business disruption and revenue.

- Estimating the cost of lost customers
- Acquiring new customers
- Reputation losses
- Business disruption and revenue losses

The IBM Security and Ponemon Institute (a security research center) conducted research on the cost of data breaches throughout the globe since 2014. Based on the 2020 report, the average total cost of a data breach in 2020 is $3.86 million, which is less than the average total cost of a data breach in 2019 by 1.5% ($3.92 million in 2019). The organizations practicing security automation and incident response processes are noted with less data breach costs, whereas the organizations that do not follow these processes are significantly higher. [8]

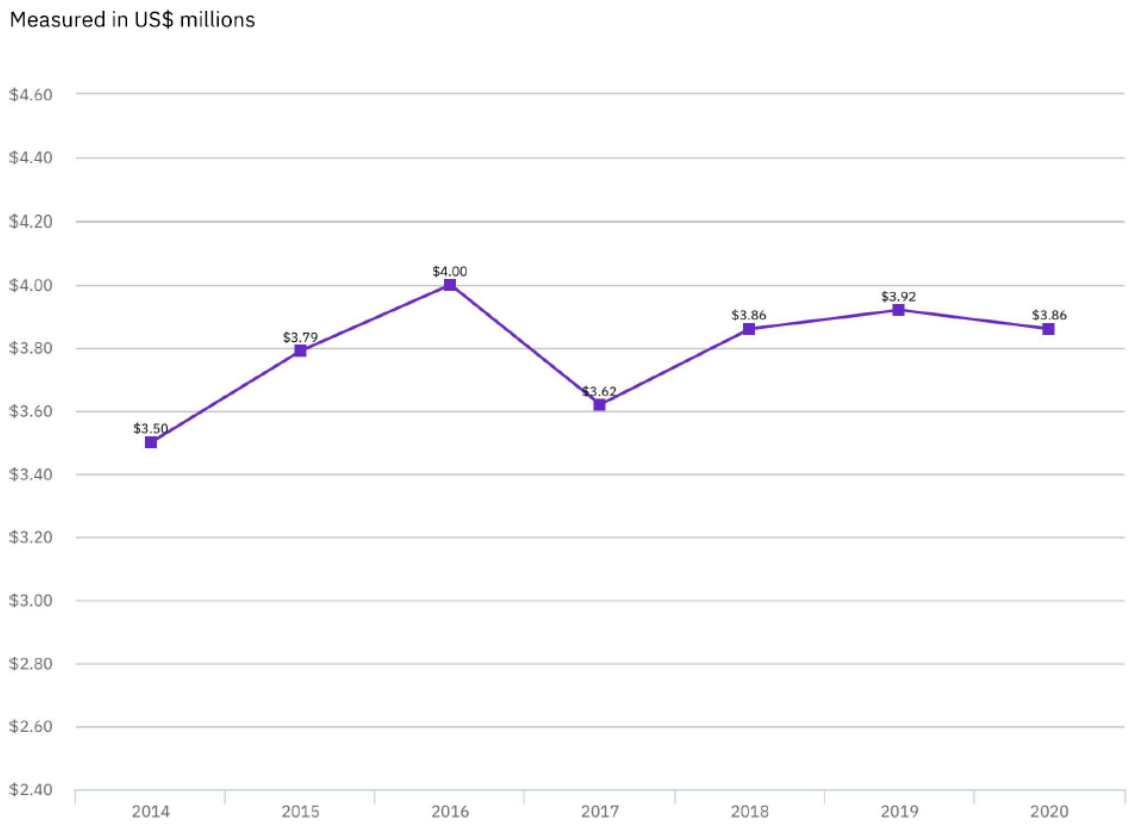

Figure 1 Average total cost of a data breach [8]

Figure 1 shows the graphical representation of a data breach's average total cost throughout the world from 2014 to 2020.

In 2014, the average total cost of a data breach stood at $3.50 million and then climbed steadily over the next three years to reach $4.00 million by 2016. The following year, 2017, it declined slightly to $3.62 million, after which it grew again to $3.92 million in 2019. Finally, it settles at $3.86 million in 2020.

Overall, the average total cost of data breaches over seven years was recorded as $3.79 million. By analysing the above graph, we can notice a 9.32% increase in the average total cost from 2014 to 2020.

In IBM's report, they have mentioned the average cost of a single lost or stolen record based on various data types. The customer's PII was the most affected and had the highest average cost of a single lost record of $150 per record. While the average cost for all the data types in 2020 was $146 per record. [8]

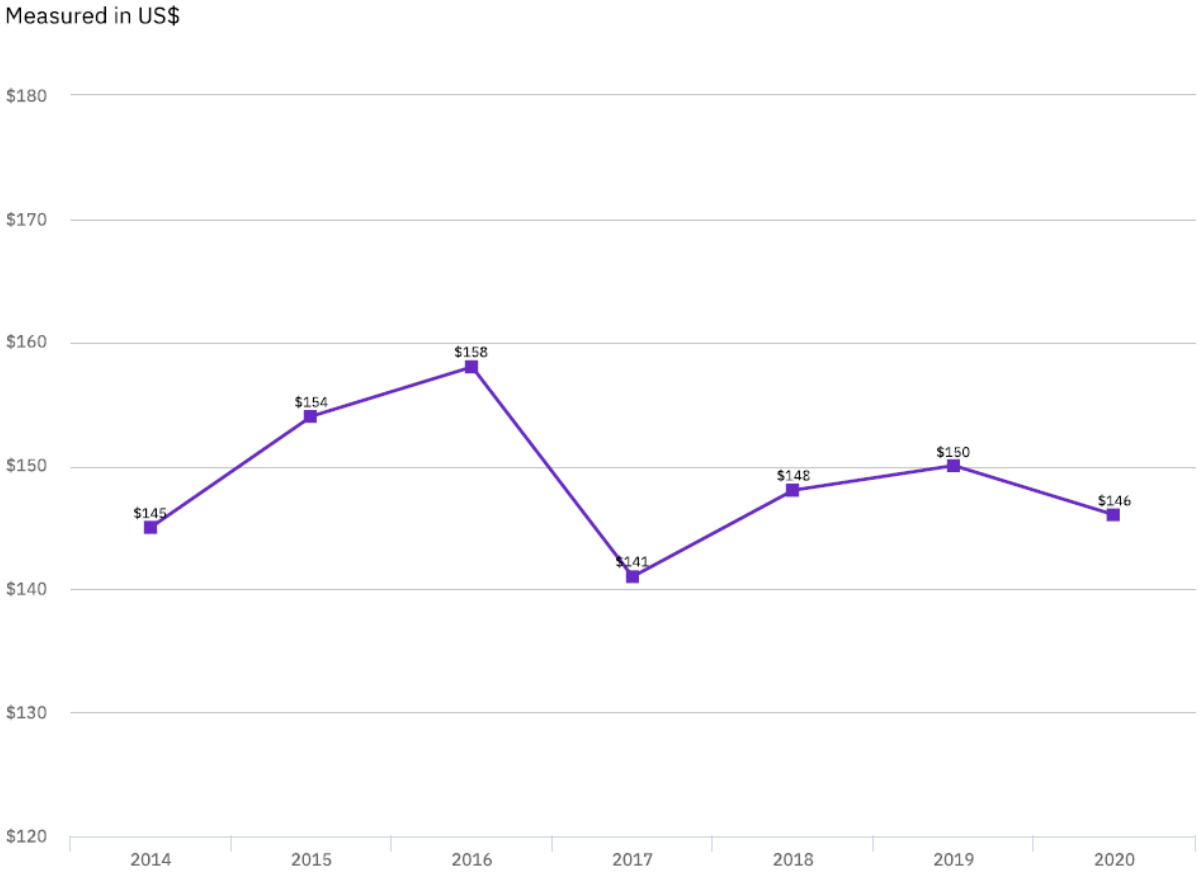

Figure 2 Average per record cost of a data breach [8]

Figure 2 displays the graph showing the average cost per stolen or lost record in all the countries from 2014 to 2020.

Starting in 2014, the average cost per record was noted as $145 and then steadily increased to the highest of $158 in 2016, subsequently reaching the lowest of $141 in 2017. Again, the per-record cost was climbed gradually for the next three years to reach $150 in 2019. Finally, after many ups and downs, it came to $146 in 2020. By examining this graph, the average cost per record over the seven years is around $149. And there was a 0.68% increase in the per-record cost since 2014.

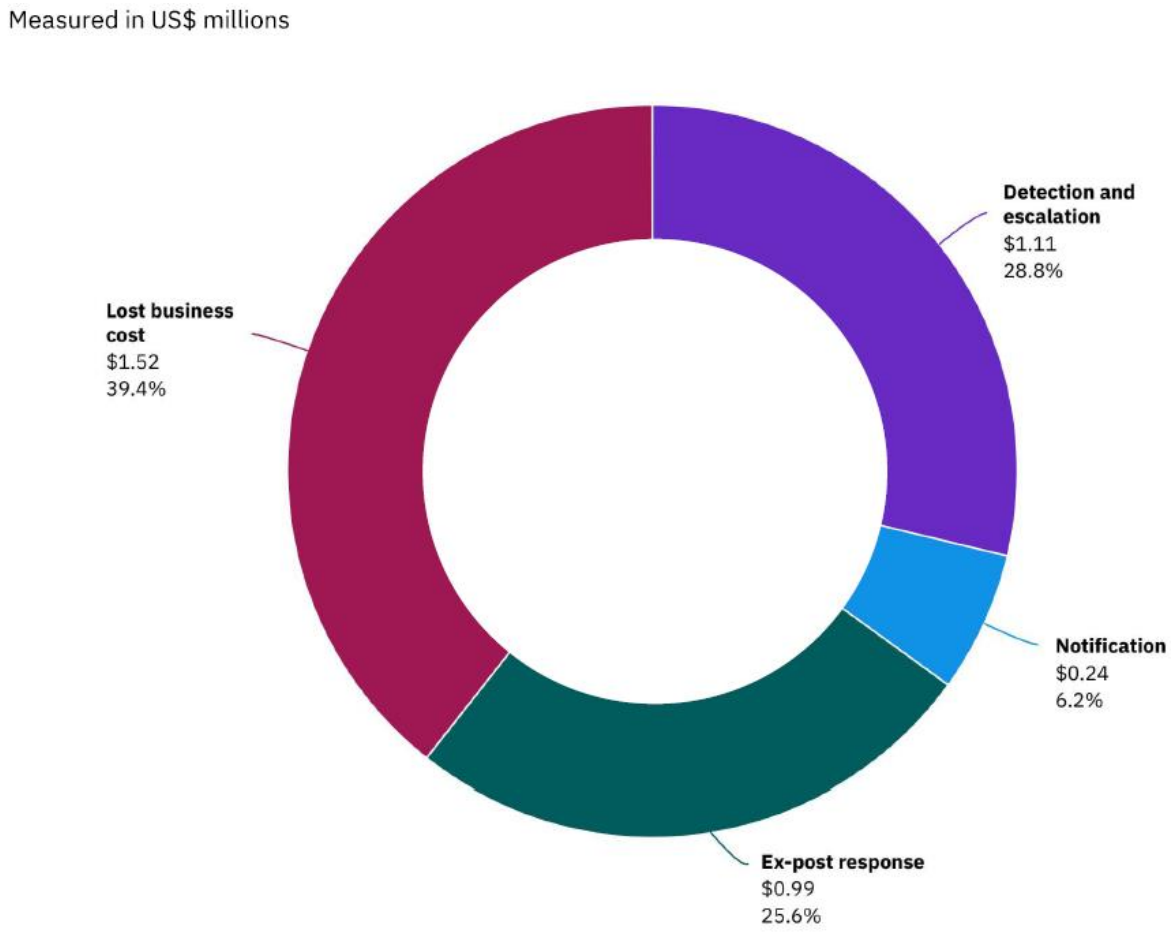

Figure 3 Average total cost of data breach in four categories [8]

Figure 3 presents the companies average total costs for four process-related activities due to data breaches throughout the countries in 2020. Loss of business was the most significant contributing cost factor, with nearly 39% of the average total cost, with $1.52 million in 2020. In contrast, the average total cost for the Detection & Escalation and Ex-post response activities were mentioned as $1.11 million and $0.99 million, respectively. The lowest was reported for performing notification activities, with $0.24 million or 6.2% of the data breach's average total cost.

Usage of artificial intelligence platforms and automated breach orchestration was rapidly grown from 15% to 21% between studies conducted in 2018 and 2020. Moreover, the adoption of security automation positively impacted the companies by reducing the costs involved in the post-incident activities. The companies which are not following security automation had an average total cost of $6.03 million, which was around twice the average cost of $2.45 million for companies that adopted security automation. Due to security automation, companies saved around $3.58 million in data breach costs. [8]

Incident Response (IR) is a primary method to reduce the average cost of the breach. The company, which had an IR team and tested an IR plan using simulation, was noted with a less average total cost of $3.29 million, compared to the companies not having an IR team and testing at $5.29 million (difference of $2 million). [8]

The average time required for the detection and containment of the data breach is based on several factors such as industry, geography and security maturity. According to the 2020 IBM study report, companies took an average of 207 days to identify and 73 days to contain the breach. The average lifecycle (sum of time taken for the identification and time taken for the containment) of the breach was 280 days. The healthcare sector's average lifecycle was 329 days, while the financial industry's lifecycle was 233 days, which was 96 days shorter than the healthcare sector. The company followed fully developed security automation had a lifecycle of 234 days based on the security standards. In contrast, the company does not follow security automation measures with a lifecycle of 308 days. This shows significant importance in implementing fully automated security measures to minimize the time. [8]

Data breaches are reported globally. Considering the costs based on the geography in 2020, the US continued at the first place in the highest average total cost of $8.64 million per data breach, followed by the middle east region with $6.52 million per breach. Furthermore, Canada stood at third position with $4.5 million, which was around twice as low based on the US. The US was in the highest position in the number of breaches because it had the highest reporting rate. For instance, most US-based companies had reported the data breach incidents that happened to them to their local authorities. [8]

Average cost of Mega Breach

A mega breach is a data breach that compromised more than 1 million records. A single mega breach can cause a devastating impact on both the company and its customers. The average cost of a mega breach is growing continuously year by year. Based on the IBM study, the

average cost of a mega-breach which consists of more than 50 million records was $392 million in 2020. [8]

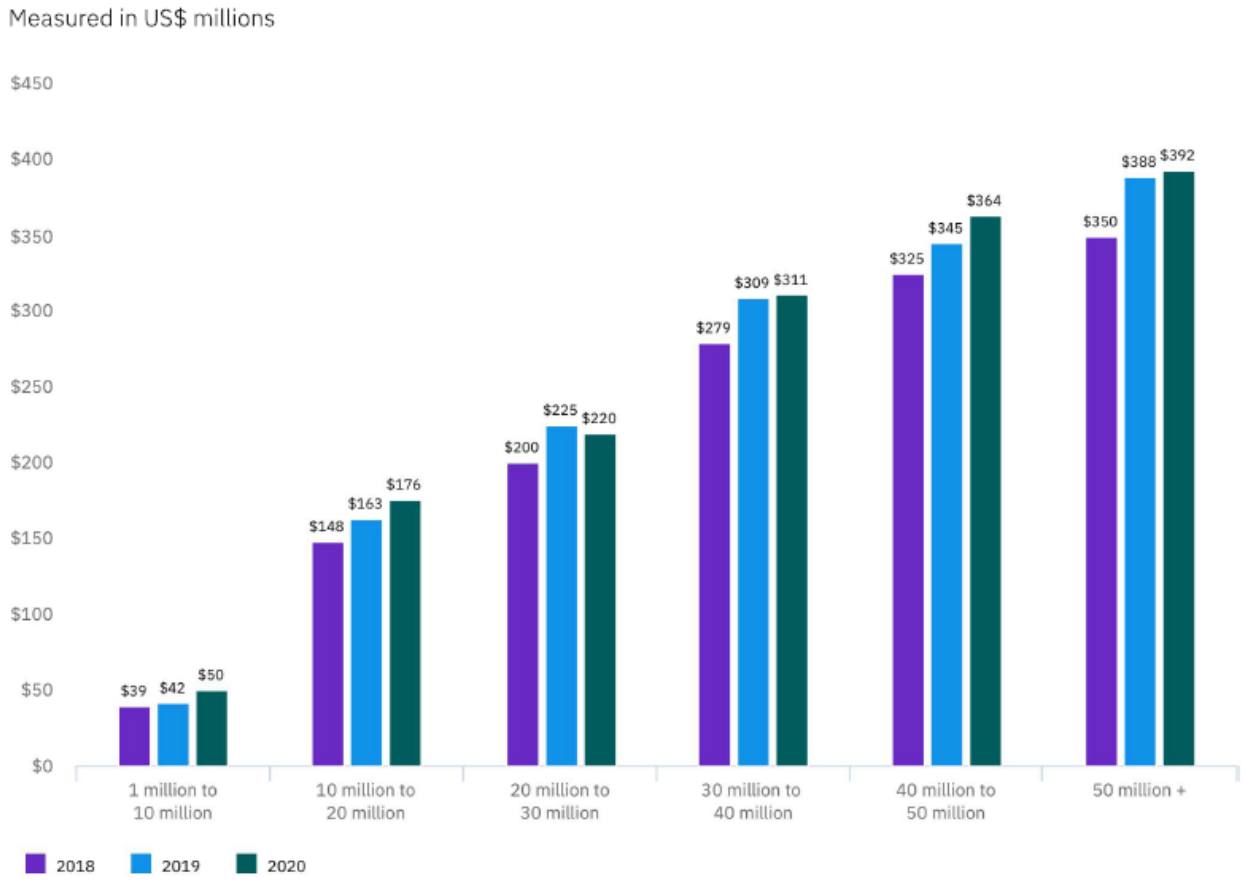

Figure 4 Average total cost of Mega Breach vs Number of records lost [8]

The above bar chart shows the average total cost of a mega-breach in 2018 and compares it with the data found in 2019 and 2020. A mega breach of 1 to 10 million records cost an average of $50 million, around 13 times more than the average cost of $3.86 million for a breach with less than 1 million records. The breach with 50 million-plus records experienced a massive growth by an increase of 10.7% from an average total cost of $350 million in 2018 to $392 million in 2020.

In 2020, the breaches with above 50 million records had an average cost of $392 million, which is around 102 times the average total cost of a data breach of below 1 million records ($3.86 million).

Chapter 3: Data Leak Threats

3.1 Classification of Data Leak Threats

Classification of data leaks can be done based on the cause such as either intentionally or inadvertently leaking the data. There is one more approach based on the parties causing the leaks such as insider threat or outsider threat. The classification is represented in a tree diagram.

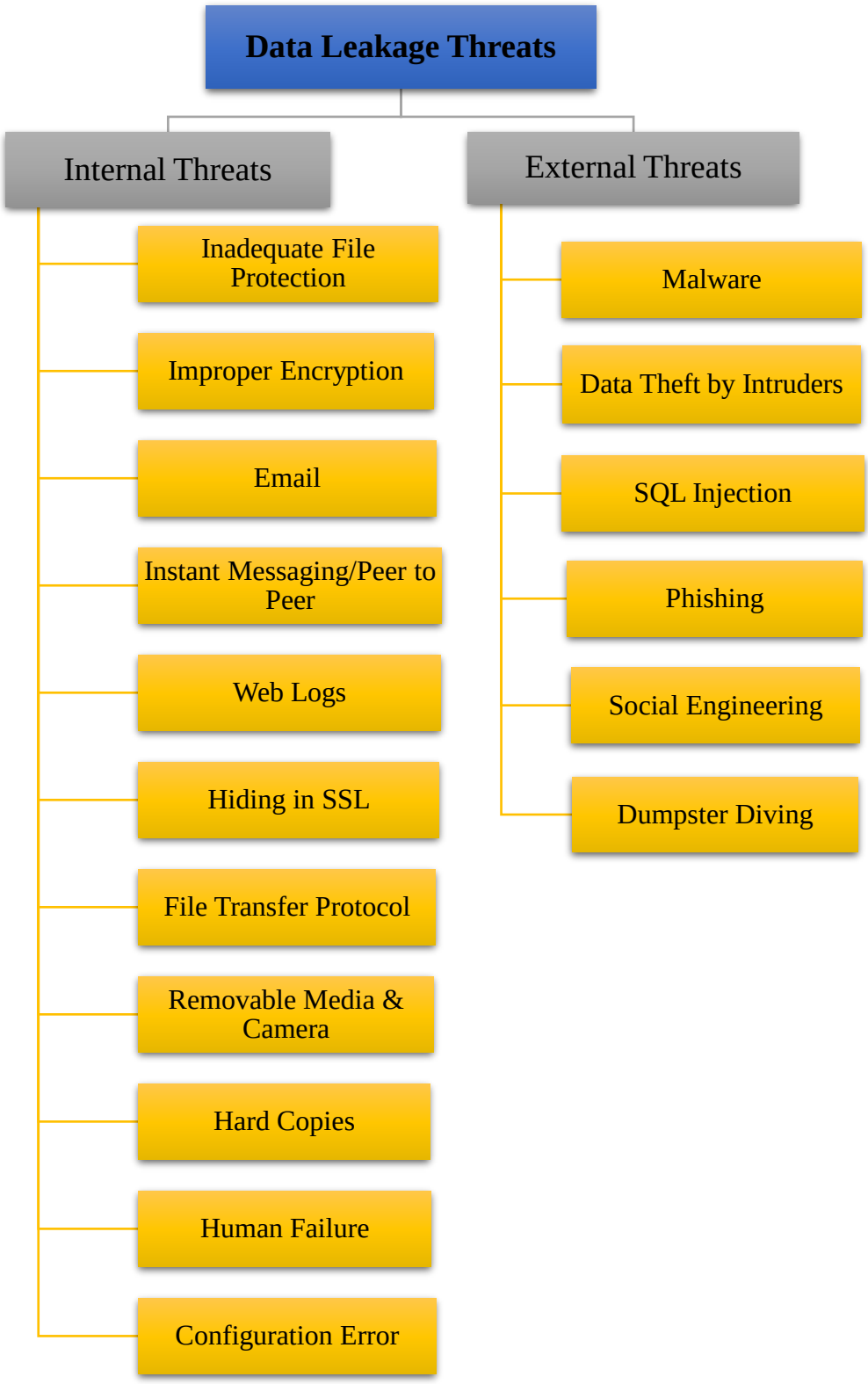

Figure 5 Classification of enterprise data leak threats [9]

3.2 Internal Threats

While the data suggests that the leading cause of internal data leakage is from inadvertent actions, there is a risk of intentional unauthorized release of the organization's data. The malicious insiders could use mediums and methods such as Remote Access, Email, Webmail, Peer-to-Peer, Instant Messaging and File Transfer protocol (FTP). They also use removable media to transfer the data. Motivations for these intentional data leaks are varied, including financial reward, corporate espionage, or a grievance with their employer. [9]

Many data leaks happen due to unintentional actions of the organization's management. This creates a challenge for businesses as the solution to this problem is not just implementing a secure content management system. Business operations should be examined and re-engineered, employees must be trained, and a cultural change may be needed in the organization. [9]

Some of the significant internal data leakage vectors are explained below:

3.2.1 Inadequate File Protection

When the files and folders do not have appropriate protection, such as user privileges and group policies, it becomes effortless for users to access and copy sensitive data from the network drive to the local system. Then the user could send it out externally using removable storage devices, email, FTP, or through other means. Furthermore, confidential data as plain text could be accessed by anyone due to a lack of encryption. [9]

3.2.2 Inadequate Database Security

Database security is a crucial aspect of every organization. Poor SQL programming could make an organization vulnerable to SQL injection attacks or allow inappropriate data to be retrieved through legitimate database queries. Moreover, without the implementation of broad database privileges such as one-size-fits-all, the confidential data could be exposed intentionally or inadvertently. The sensitive data in the database, such as passwords, should be encrypted because password leakage will cause devastating damage to the organization. [9]

3.2.3 Email

Many organizations use traditional email clients such as Microsoft Outlook, an internal user could send an email with a confidential document to an unauthorized person. They may also use compression techniques and encryption techniques to disguise the presence of confidential

data. Steganography may also be used to embed the confidential data within normal data stream. Alternatively, the document data should be copied into the email body instead of attaching the entire document. [9]

Email could also act as a vector for inadvertent disclosure. An employee could attach the wrong file inadvertently or may select the wrong recipient. In some cases, employees may be tricked into sending a document through social engineering. [9]

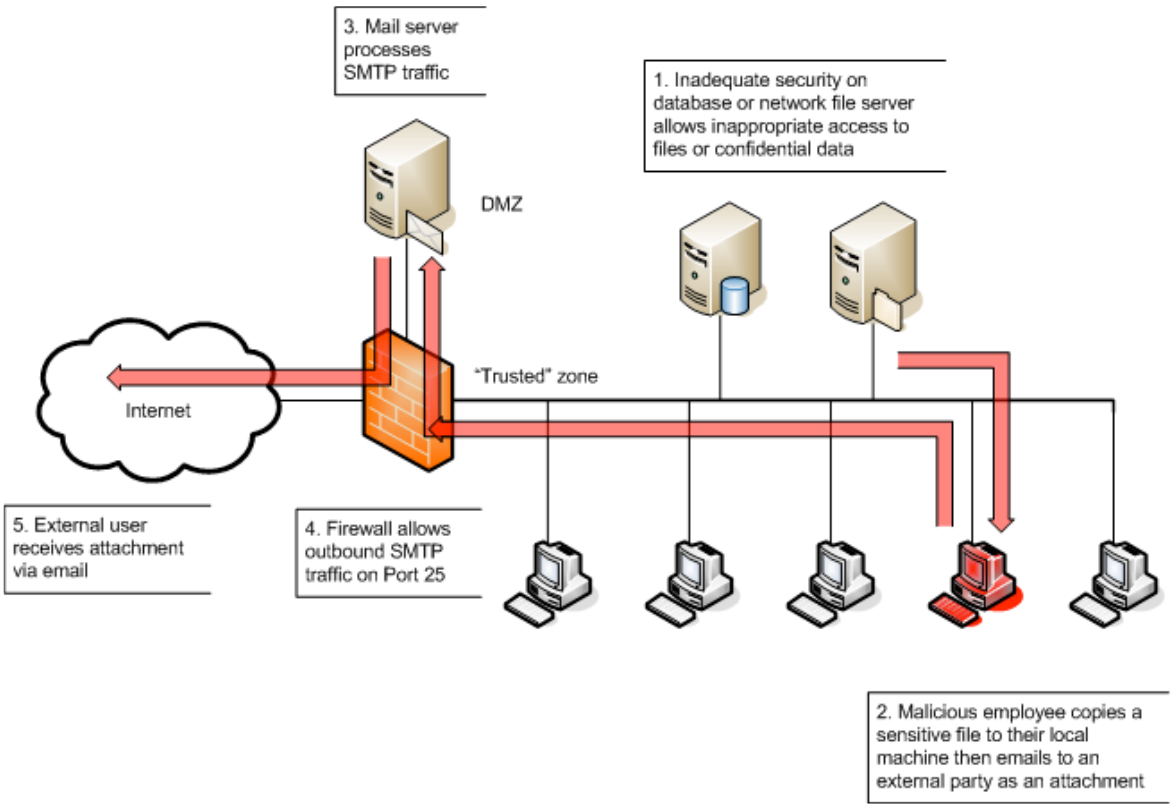

Figure 6 Email Data Leakage Vector [9]

3.2.4 Instant Messaging and P2P

Several organizations allow their employees to access instant messaging services such as Skype, Google Hangouts, Yahoo Messenger from their workstations. Many of these services can transfer files enabling their employees to send sensitive documents to an unauthorized person. Moreover, a user could disclose confidential data through chat sessions. [9]

Instant messaging is also becoming a vector for the malware distribution. The following figure explains how instant messaging causes the data leak. [9]

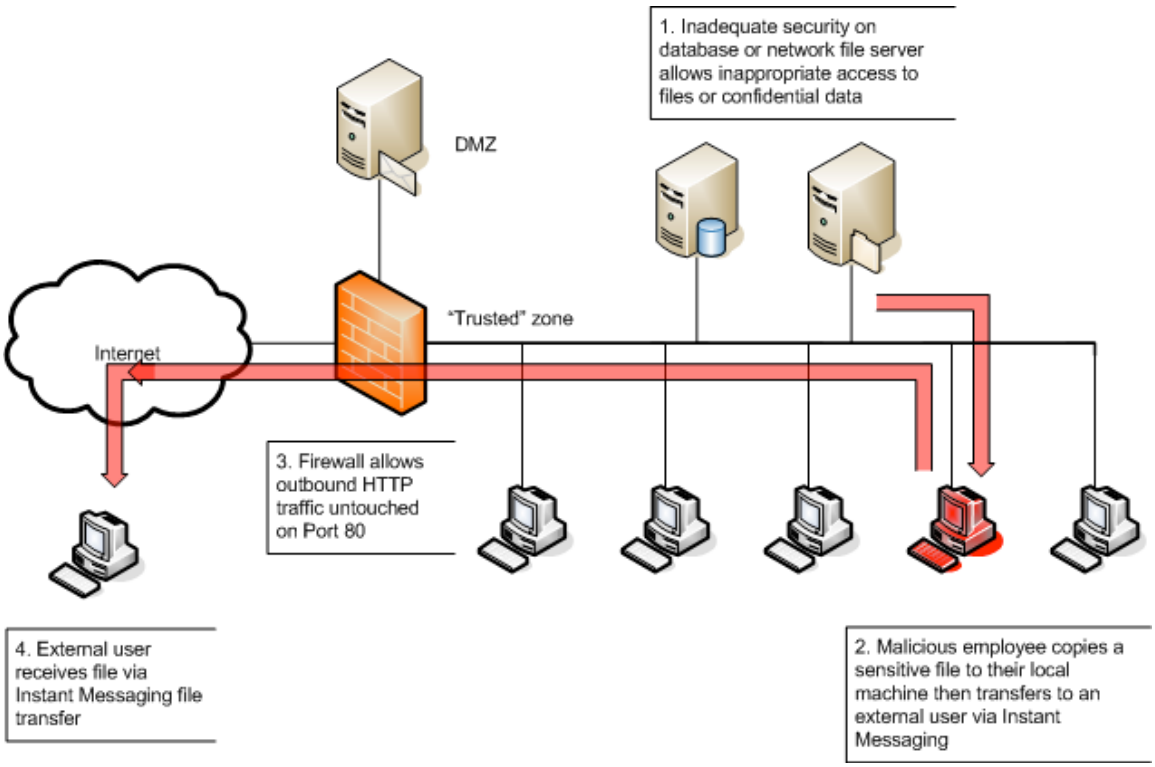

Figure 7 Instant Messaging Data Leakage Vector [9]

Peer-to-Peer applications allow users to share the data in different formats over the internet. Basically, a single server provides the data to the end-users, but in the P2P network, any computer in the network can act as the server and serve the peer, acting as a client. P2P is one of the significant threats to data confidentiality, and it can lead to data leaks. Users may be unaware of what data they are sharing and unaware of data the application download and store on their systems, making them host inappropriate data. [9]

3.2.5 Web Logs

Web Logs, also known as Blogs, are the websites where users can write their content, comments and opinions on a particular topic. These sites can be owned by them or maybe public sites that could contain input from many individuals. An employee from an organization could post confidential data in his blog intentionally or inadvertently. However, he could be tracked easily, so this method is less likely used. [9]

3.2.6 Hiding in SSL

Users could utilize a public proxy service via SSL connection known as proxy avoidance to obfuscate the data. They can assess the proxy using a web browser and type the URL of the site they want to visit, then the entire session with that site should be encrypted. A firewall with

stateful packet inspection will not inspect the encrypted data, so a malicious user could leak the sensitive data. [9]

3.2.7 File Transfer Protocol

FTP is also considered as an internal threat vector for the data leak. It is easy to install and configure a basic FTP server external to the company. The individual then installs a publicly available FTP client and upload that file to the server. This could use a "dead drop" public FTP server hosted offshore, where third-party users have access to it. As FTP is a popular protocol, it is likely to be allowed through the firewall.

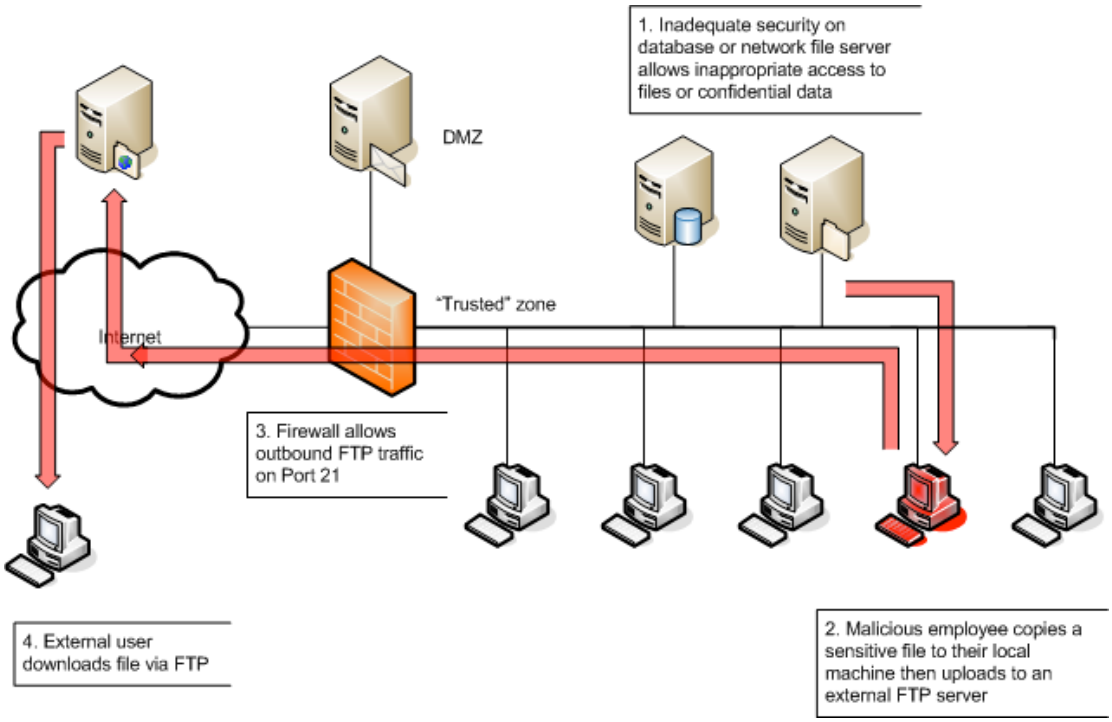

Figure 8 FTP Data Leakage Vector [9]

FTP is mostly used for intentional leakage rather than unintentional, because uploading a file to the FTP server is not performed by regular users on a daily basis. [9]

3.2.8 Removable Media and Cameras

Based on the Symantec Internet security threat report, theft or loss of a computer or data storage medium was the reason for most of the data leaks, up to 57% of all identity thefts. [9]

Nowadays, removable storage devices such as flash drives and pen drives are cheaply available. Copying a large document in any format onto the flash drive is effortless, and it also take less time due to high data transfer speeds making an attacker to steal the data easily.

The Universal Serial Bus (USB) keys are small in size, so, they are more likely to lose. Although copying of data is legitimate, there is a risk exists that any third party could find the drive after losing it.

A motivated individual may take the digital copies of the screens or any documents. A camera is not needed nowadays, most of the mobile phones are coming with a high megapixel built-in camera. The taken photos can be shared quickly to the cloud or to another person. [9]

3.2.9 Hard Copy

Suppose an attacker wants to provide sensitive documents to the competitor of a company. In this case, it is possible to print the data into hardcopy and walk out of the office with their briefcase despite the victim company implemented electronic countermeasures. Alternatively, they may simply place it in the envelope and mail it. [9]

3.2.10 Human Failure

Human failure is one of the main reasons for the data leaks, and it comes under an internal data leakage threat. Human errors are more common and frustrating, because these can be avoidable. An employee in an organization may make mistakes, such as clicking on a malicious email that could infect the system with malware which could cause a data leak.

However, a dissatisfied employee in an organization could disclose the sensitive data intentionally due to many reasons such as financial gain and personal revenge on their superiors. The employee is no needed to be a long-term employee but can also be a contract employee who can access the company resources.

3.2.11 Configuration Error

The configuration should be performed correctly. Any configuration mistakes could cause vulnerabilities in the system, finally lead to the data leak. At some point, the configuration error may arise, but implementing a fail-safe mode to prevent data loss is the best practice. [10]

3.3 External Threats

Any threats from an external source or an outsider with malicious intent are known as external threats. These threat vectors from outside could cause a devastating impact on the organizations both financially and reputationally.

Some of the external threats are explained in the following sections.

3.3.1 Malware

Day by day, attackers are creating several new malwares and introducing them to the websites or embedding within the software applications. After infecting a system, a malware can scan through the folders and sends out randomly selected documents or any selective documents containing sensitive data to the attacker.

If the malware is a zero-day threat, the malware's signature is not available, so there should be a higher risk of bypassing the inbound gateway protection measures and even anti-virus installed on the system. Once the system is affected by this malware, outbound communication is established and sends sensitive information outside. A typical firewall could not restrict the traffic, which is going out from the internal device, making the data leak effortlessly. [9]

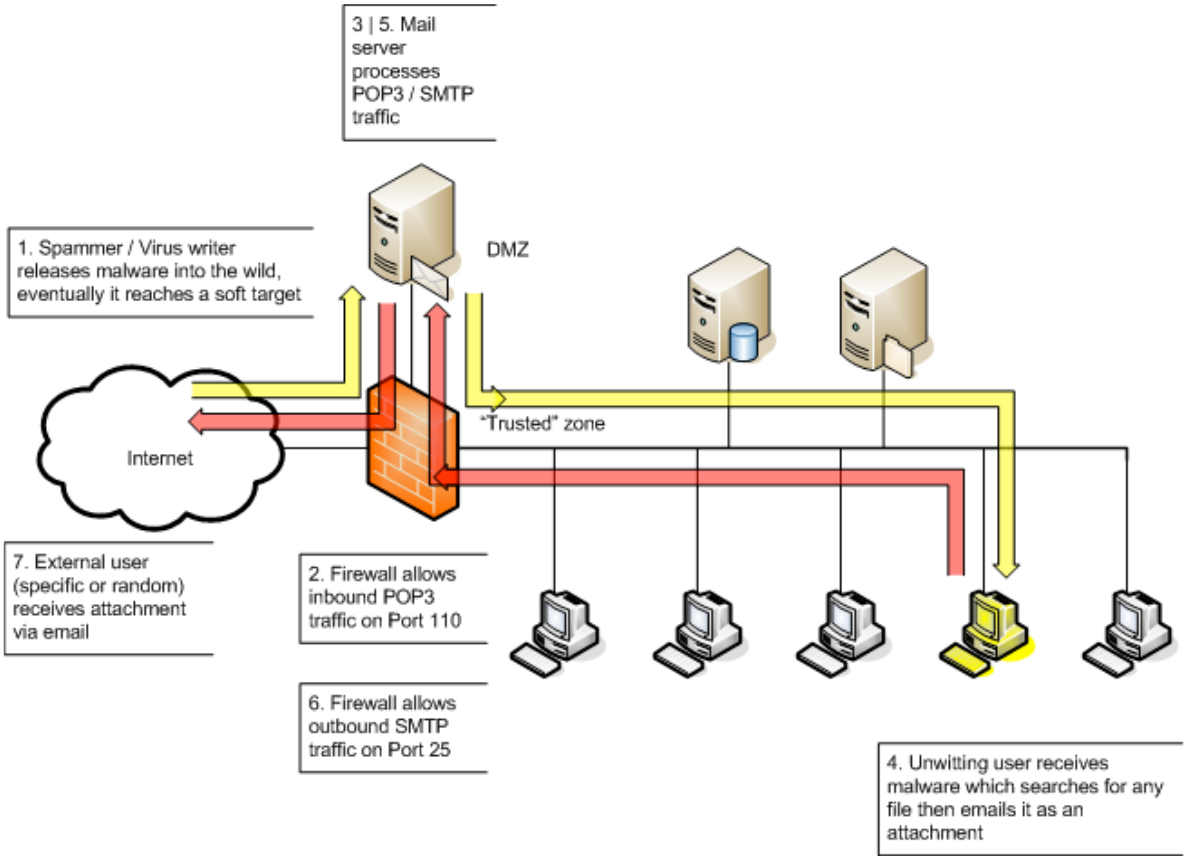

Figure 9 Malware Data Leak Vector [9]

Figure 9 describes how malware bypasses the firewall and sends private data outside the network. Besides malware, a keylogger could capture all the keystrokes entered by the user, including login credentials and sensitive data, which results in identity theft or even a data leak. [9]

3.3.2 Data Theft by Intruders

Electronic break-in to an organization happens very often. The intruders break the security systems and steal sensitive information such as credit card information. Later this information will be placed on the dark web for sale. Based on the Monster.com case study, hackers hacked the website, collected all the resumes on the site, and used these documents to create personalized phishing emails to the job seekers. This could be concerning because the hackers may also create a fake website and make the users send their resumes directly to them. [9]

3.3.3 SQL Injection

In an SQL injection attack, SQL queries are used to retrieve unauthorized data from a database. Websites using an SQL Server as a backend database is vulnerable to these attacks. This attack is most possibly due to a lack of proper input validation techniques used in a login form. Additional data is entered within the input data on the web page form to generate different SQL statements. For example, an attacker could enter extra information such as [11]

Abhijith'; SELECT * FROM Students; --

Here semicolon indicates end of the query, and two dashes indicate the ignored comments. If the website took this string directly into the SELECT statement surrounded by the same single quotes, it looks like this: [11]

SELECT * FROM Students WHERE Name='Abhijith';
SELECT * FROM Students;

The first query retrieves the data related to the Student name 'Abhijith' as usual. However, the semicolon indicates the end of the query and the second line is taken as a new query and retrieves all the data from the Student table. Like this, an attacker could get sensitive and unauthorized data from the database. [11]

3.3.4 Phishing

Phishing is one of the main reasons for the data leak. Phishing is an activity in which users are sent emails to trick them into clicking on the email's links to reveal their personal information or direct them to the malicious website that appears to be an original site. For example, a user received a phishing email that looks like it came from a well-known company.

In reality, the user even doesn't have an account in that company. However, some users would click the link in the email curiously. When they click, they might be redirected to the phishing website and asked to login into their accounts with their credentials. After entering credentials

and tried to log in, they again turned to the original website, and the attacker would access this username and passwords. It is also possible to direct the users to the malicious websites that host malware and make them download the malware so that the attacker could gain remote access into the user system.

3.3.5 Social Engineering

Social engineering uses social tactics to manipulate people and gain sensitive information. It is often focused on the individuals, encouraging them to perform some tasks or reveal sensitive information such as credentials. Techniques used in social engineering are: [11]

- Encouraging someone to perform a risky action
- Encouraging someone to reveal sensitive information
- Impersonating someone as an authorized employee
- Flattery and Conning; learning secrets about different professions by conning and flattering people into telling them. And using the learned secrets to impersonate them and then perform some illegal activities.

The information such as name, address, educational and professional details could be used to answer the login process's security questions. So, if a social engineer succeeds in getting this information, he can use it to reset the user account password and then finally gain unauthorized access to the account.

3.3.6 Dumpster Diving

Some organizations do not follow proper disposal methods to dispose of their hard copy records containing confidential information. Due to this, a person with malicious intent could raid the company's dumpster and obtain this information, which could profoundly affect its business. Moreover, electronic media such as hard disks, CDs and DVDs contains the company's confidential information. So, proper disposal techniques such as shredding, wiping, erasing, burning, and degaussing are essential to prevent these attacks.

Chapter 4: Data Leak Prevention

Generally, companies handle several types of data such as employees, customers and business transactions. This data has threats in the form of data loss and data leakage. So, data loss prevention techniques are developed to prevent data loss and data leakage. Commonly data loss prevention and data leakage prevention are used interchangeably, but these are two different terms. Data loss prevention (DLP) techniques can prevent the loss of data during general use. In contrast, data leakage prevention techniques prevent the transmitting of data outside of the organization. [12]

Data loss prevention methods such as backup files and anti-malware software prevent data loss due to hardware failures, natural disasters and malware. Simultaneously, Data leakage prevention techniques are used to implement privacy laws and data flow maps to safeguard crucial data within the organization. [12] However, the commercial Data loss prevention solutions have the ability to handle both data loss prevention and data leakage prevention.

To understand the DLP solutions, we need to know some basic terms related to the DLP. Some of them are explained below:

4.1 Data State

The data in an organization should be in any one of three phases throughout the lifecycle of the data. The three phases are Data-At-Rest, Data-In-Use and Data-In-Motion.

- **Data-At-Rest:** The data stored in a local hard disk on a computer or remote storage such as cloud is known as Data-At-Rest. The protection of this data could be achieved by using security measures such as encryption and access controls.
- **Data-In-Use:** When the user is using or interacting with any data, then this data is said to be Data-In-Use. To protect and monitor this data, endpoint-related systems are implemented.
- **Data-In-Motion:** The data in the transmission on a network is known as Data-In-Motion. This data could be inbound or outbound traffic data coming into the internal network or going out from the internal network. The DLP techniques such as packet inspection can detect the type of protocols used in the network transition.

4.2 Leakage Handling Approaches

There are mainly two approaches for handling the data leakage. They are:

4.2.1 Detective Approach

In this approach, the data leakage incidents are detected by the DLP system, and it will take necessary corrective actions to handle the data leakage. For example, if any file with confidential data is transferred to an external server, the DLP agent detects and tries to stop it from sending out of the network. This approach has three forms, they are: [13]

Context-based Inspection

In this type of inspection, contextual information such as source, destination, sender, recipient, size, metadata, time stamps, location, format, application and transactions are extracted from the monitored data. [13] For example, when a file is sent through a network, a network based DLP can detect the transmission and it extracts the contextual information of the file.

Content-based Inspection

In this approach, the detection of data leakage is done by analysing content. There are several techniques for content inspection. [13]

- A combination of lexicons containing patterns and keywords such as “confidential,” “financial-report” is attached to every document to identify different type of data. Most of the products have lexicons that address laws and regulations. Based on these keywords, the detection of data leakage becomes easy and fast.
- A fingerprint is a unique hash value for a set of stored data in the database. The fingerprints of sensitive files are collected and searched for exact fingerprints of files in the data leakage by comparing the hash value with the inspected data.
- In the natural-language analysis, sensitive data and inspected data are compared to check whether they are similar or not.
- Statistical data is extracted from the content obtained from the inspection by using machine learning methods. This technique is efficient for unstructured data.

Content Tagging

A tag is given to every file containing confidential data, and then a policy is enforced based on these tags. This tag remains to the content even while an application is processing this data. The tags can be assigned manually while creating the file or automatically to a specific group of files by using context-based analysis. [13]

4.2.2 Preventive Approach

The preventive approach name itself shows that the data leakages are prevented by using prevention measures. DLP uses several preventive approaches such as:

Access control

While accessing any resources, DLP can permit or deny based on a particular entity. If access to the confidential information is granted, a DLP can restrict information based on the policy. The best way to automatically provide access control to any document is by combining it with Enterprise Digital Rights Management (EDRM). [13]

Disabling functions

Some functionalities can be disabled to stop inappropriate access to confidential files. For example, a copy-paste function could be disabled for some files due to the company's policy, which restricts the employees from copying confidential data into a portable storage device such as a USB flash drive. [13]

Encryption

Encrypting the data is one of the best practices to restrict confidential data access by any unauthorized entity. A company could implement this policy, stating what type of data is encrypted and who can decrypt the data.

Awareness

Every organization should provide proper training to their employees about sensitive data, who can access it and how to be careful while accessing it. Awareness could be given by conducting awareness camps, training sessions.

The preventive measures are discussed later in detail in this chapter.

4.3 DLP Policy

A Data Loss Prevention policy (DLP Policy) consists of conditions, exceptions and actions used to evaluate and define the scope of data, files and documents to detect and prevent data leakage. Every DLP tool is configured using a technical DLP policy, which is different from organizational policy. Based on this policy, organizations can define: [14]

- What data can and cannot be transmitted
- Where data can be transmitted
- How data can be transmitted
- Who can send and receive data

Based on these conditions, a DLP tool can decide on what type of data it should monitor and what type of action it should take. If any data matches with these conditions, the system sends a policy violation generated by a user or system, known as 'Incident'. When a policy violation occurs, actions stipulate how the DLP tool should act to handle the incident. Several types of actions can be applied depending on the number of incidents and the risk level. While checking the conditions, any activity or data can be exempted from matching the condition based on the Exception rule. [14]

After creating a DLP policy, it can be applied to one or more data leakage channels. Moreover, there is no need to apply to the whole enterprise; it can be applied to a group of users or geographic region. [14] The table shows examples of the DLP policies.

DLP POLICY	CONDITION	EXCEPTION	ACTION
Confidential File	Detect content that matches keyword 'Confidential'	No Exception	Block transfers containing confidential files
Project ABC	Detect content that matches 'Project ABC' and intended for external recipient	Allow transfer to External Legal Counsellor	Block transfer to external recipient Encrypt data sending to the External Legal Counsellor
Medical record	Detect content that matches words from a medical terms list.	Allow personal emails	Block copying of medical record data into portable storage device. Display notification stating that a file transfer violates a DLP policy.

Table 1 DLP Policies [14]

A DLP policy can be implemented for three reasons. They are: [15]

- **Compliance**: In every country, the government impose some mandatory compliance standards on the organizations to maintain data privacy and security. For example, in Canada, every organization which handles personal and health data should follow the standard 'Personal Information Protection and Electronic Documents Act' (PIPEDA). A DLP policy is used to implement these compliance standards.
- **Intellectual property**: A DLP policy helps identify sensitive data and safeguard the critical information assets in an organization.

- **Data Visibility**: A DLP policy can provide a clear view of data to its stakeholders in a company.

The DLP policy can be created by using predefined templates or building custom policies. Based on the organizational needs, these templates can be customized. Most DLP tools provide a library containing predefined policy templates to detect the type of data based on the regulatory requirements implemented by the local governments in those countries. Each template supports industry-specific data using a standard format and country-specific data such as social insurance numbers in Canada.

While creating a policy, a DLP template uses data identifiers and logical operators such as AND, OR, EXCEPT to generate condition statements. All the files come under a DLP policy after they meet certain condition statements. [15]

4.4 Process of Data Leakage Prevention

The main purpose of the Data Leakage Prevention (DLP) is to detect and prevent the unauthorized disclosure of sensitive data and also to prevent mishandling of data. In order to achieve data protection, DLP has to follow three core activities. They are:

4.4.1 Identifying Data

Any DLP should protect sensitive data instead of protecting entire data in an organization. Because the protection of entire data requires many resources, it also becomes an intensive task, resulting in the server's degradation and a decrease in network performance. Hence, identifying sensitive data is an essential task for the DLP solution.

As we discussed earlier in Chapter 2, "Analysis of Data Leak," there are several types of sensitive data such as trade secrets, PII, PHI and credit card information handling by the organizations.

A DLP tool first task is to identify what type of data needed to be included in the DLP scope. A DLP tool should be appropriately configured to avoid configuration mistakes resulting in the false detection of the sensitive data, making the sensitive data unprotected during data leaks.

A DLP tool uses data detection techniques for identifying sensitive data. The detection is done by searching for a predefined keyword or any alphanumeric patterns within the data files. The detection process may fail in some cases because the DLP tool cannot access specific data files due to data obscurity in the form of encryption and compression.

Some of the data detection techniques used by the DLP tools are described previously in Section 4.3, “Leakage handling approaches.” However, for clarification, those techniques are described below. [14] [16]

- **Described Content matching**: Detection of data is done by matching the content with keywords, patterns, regular expressions, or dictionaries contains a list of specific terms. For example, a particular type of keywords such as ‘confidential,’ ‘private’ and ‘restricted’ are used as tags to the files to describe the file’s scope. By using these tags, a DLP tool can detect it and prevent it from unauthorized transmission.
- **Fingerprinting**: A fingerprint is a cryptographic hash value of a sample file. A DLP tool can check the file’s content with the fingerprint to detect the complete or partial match. This technique has a drawback; a file is required to generate a fingerprint before performing actual verification of the content.
- **Database Fingerprinting**: It is also known as Exact Data Matching. Data is verified with database dump or live database to get an exact match. This technique may cause performance issues due to live database connections.
- **Machine Learning**: Some example documents are provided to the DLP solutions to understand the type of data in the document, then algorithms and statistical techniques are used to verify the content with example documents to determine it. This technique is used for the data such as source code and software design documents, which is not possible to get the fingerprint of it.
- **Optical Character Recognition**: This technique is also known as Image recognition. In some cases, data can be exfiltrated by taking screenshots and scanned documents. In such cases, image recognition analysis is helpful to detect the sensitive data by extracting text from the image and matches with sensitive data.
- **Statistical Analysis**: Machine learning methods or statistical methods such as Bayesian analysis is used to obtain statistical data and to trigger any policy violation in the secure content. A large volume of data is required in this method because fewer data samples lead to false positives and false negatives. Unstructured data can be identified effectively by using this technique.
- **Pre-build Categories:** Pre-build categories are defined, containing rules and dictionaries for sensitive data such as Payment Card Industry (PCI) protection. The data is then verified with these categories. If any data is matched, then the data should not be allowed to the transmission.

4.4.2 Monitoring Channels

As we discussed in Chapter 3, 'Data Leak Threats,' there are many channels through which the data is exfiltrated, such as email, webmail, instant messaging, weblogs, and FTP. Furthermore, the data could be exposed through these channels intentionally or unintentionally due to threats from both inside and outside the organization. Potential leaks can be detected by monitoring these channels, and risks associated with the data can be understood.

A standard DLP tool can identify all the channels through which data leaks might occur. However, it cannot monitor all the channels due to resource constrain, financial costs and even the system performance. This DLP tool encourage the organization to prioritise the channels based on the leakage occurrences. The most causing data leakage channels are monitored continuously whereas the remaining channels are monitored less. [14]

DLP tools use different monitoring techniques based on the state of the data. As we discussed earlier in this chapter, the data can be in any one of three types of states: Data in motion, Data in use and Data in rest. Monitoring techniques for these data are described below: [14]

- **Data in Motion:** Network DLP tools monitor the network by sniffing the packets transmitting on the network. These data packets are also examined using deep content inspection to identify the sensitive data sending through various channels in the network, such as email and file transfer.
- **Data in Use:** Endpoint DLP tools install some agents on the devices to monitor the user actions and even block them. A DLP can show how users are interacting with the data on the endpoint system by installing these agents. For example, a DLP agent can scan the data in use and blocks the unauthorized transfer or encrypt the confidential data.
- **Data in Storage:** Data can be stored in various locations such as repositories, systems and the cloud. By using DLP tools, this data can be inspected. This inspection process is conducted by indexing, opening, reading and analysing the data in files to detect the sensitive data. After detection, this data is moved to a secure location or encrypt the data in plain text format. The scanning operation can be performed remotely or locally in real-time, or at regular intervals or when required.

4.4.3 Act to Prevent Data Leakage

After successfully identifying and monitoring data channels, a DLP is responsible for preventing data leaks by intervening use and transmission of data to reduce the risk associated

with the low secured business processes. This can be done by handling the data by blocking it to prevent data exfiltration. The prevention can be achieved by the actions taken by the DLP tools and also the physical security measures taken by the company's management.

If any policy violation occurs, a DLP tool will act in three ways, also known as modes. They are Log, Notify and Block modes. These modes are essential to implement the DLP policies appropriately. [14]

- **Log**: Every activity, including policy violation, should be recorded and stored in the form of log files. Later, the security team uses these log files to analyse and investigate what exactly caused the data leak.
- **Notify:** Once the DLP policy rule is created and implemented, a notification is displayed to the individuals who are making the violation when a policy violation occurs. Even the violation happens after notifying the user, then a copy of the message should be sent to the management. The notification process can be done by sending an email to the user mentioning that their behaviour breached the corporate policy but still allowing the activity by guiding with the correct way of handling data. There is another way of notifying, where the user will get a pop-up warning message to cancel the data transfer. [14]
- **Block:** Blocking is the third type of action taken for handling the policy violation. A DLP tool decides to block any activities based on the activity threat level after notifying the user. Blocking is of three types. They are [14]
    1. **Hard Block:** This type of block will strictly prevent the actions from happening. For example, blocking the email containing sensitive information, disabling a user to download, copy or print any document and deleting attachment or file during transmission.
    2. **Soft Block:** Blocking the activities by showing or performing alternative ways such as moving a file to another location when it detects the file is in an insecure location. Redact sensitive data in web posts or email and allow the transfer.
    3. **Other actions:** The actions are taken to remediate the inappropriate handling of data. For example, changing the access controls to restrict access to the file by an unauthorized person and encrypting it when it detects a transmission without encryption.

4.5 DLP Architecture

Technical architecture is required to deploy a DLP tool. This architecture should be designed carefully to integrate with the infrastructure which is already existed in the organization. The new implementations should not affect the existing infrastructure negatively. A small mistake in the architecture design would affect the company's business. A DLP architecture has four components, which are mentioned in the following figure 10.

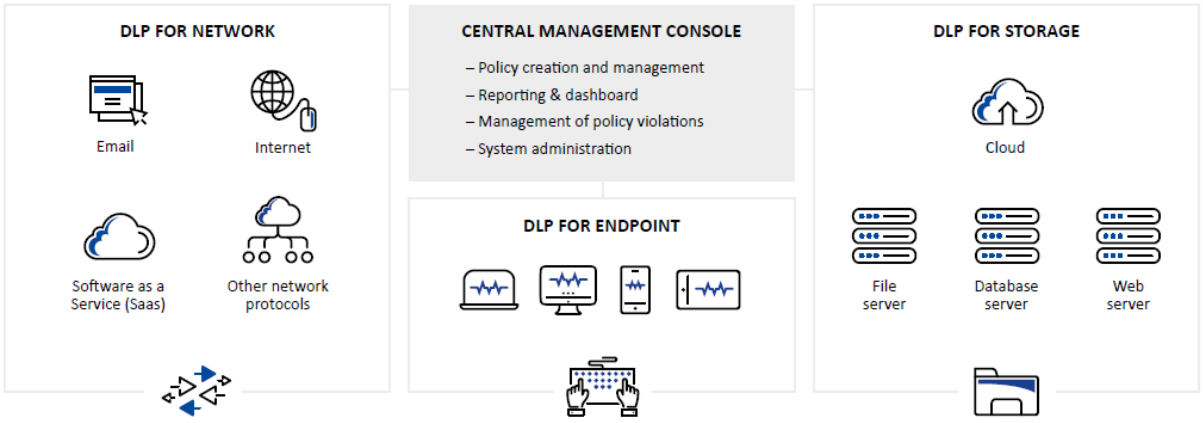

Figure 10 DLP Architecture [14]

The central management control is an essential component of the DLP because it provides a single interface to the security team to implement and manage its DLP policies. This component is also responsible for collecting all the logs from various DLP components and reporting them to the user. It is also helpful in incident repose tasks and for the system administration. The remaining DLP components identify and monitor the data in the storage, data in the usage in the systems and data in the network. [14]

4.6 Limitations of DLP

Although using DLP provides many benefits, it also has some limitations due to several factors. A DLP itself cannot prevent data leaks in all the channels due to gaps and capabilities. The limitations of DLP are mentioned below: [14]

- A DLP cannot detect all the sensitive data. As we know, data is scattered across different environments, and this data also replicated in several storage repositories and platforms such as cloud services and data centres located in different locations. In some cases, the data might be accessed from the outside of the company's network, which cannot control by the organization's DLP. So, a typical DLP cannot scan and protect this data from data leaks.

- DLP cannot monitor all the data leak channels. For example, a DLP tool does not prevent data leaking in the digital format, such as printed documents or audio format, i.e. a malicious insider could send confidential information through a phone call. Furthermore, a DLP can also be useless in situations like lost or stolen devices.
- The DLP controls may be bypassed. When a DLP focusing on a few selected channels, the malicious insider can bypass the controls and transfer the confidential data through less secure channels.
- If the DLP tool is not configured correctly, there may be a chance to allow the data leakage. When the DLP policies are defined too specific, it leads to false-negative alarms (failure to detect data leaks). When these are defined broadly, it may cause false positive alarms (detection of policy violations that are not intended to match).

4.7 DLP Solutions

Nowadays, there are many DLP solutions available in the market. However, these are not meant for all types of business needs. Before selecting a DLP solution, the following terms should be considered:

- **Network DLP:** A Network DLP software can analyse the traffic on the network, and an event can be raised if any suspicious traffic is detected, then that traffic is blocked. It maintains a database about the data being transmitted and who is sending the data. Moreover, it provides the visibility of the data in transit on the network.
- **Storage DLP:** This DLP can identify and monitor sensitive data in storage devices. It provides information above the files in storage devices and across the organization. It also provides visibility of the data stored in the local storage devices and cloud storage.
- **Endpoint DLP:** An Endpoint DLP software identifies and monitors the data on end devices such as workstations, servers and mobile devices. A remote supervisory server controls the administrative tasks, policy distribution, and log events, while a DLP agent installed on the device controls and monitors the data. This software provides visibility of the data on end devices stored locally or remotely.
- **Cloud-Based DLP:** This DLP protects an organization's data stored in the cloud by implementing cloud data policies. When an organization uses cloud storage, this DLP monitors for any sensitive data like PII and sends an alert and blocks PII transmission into the cloud storage.

Some of the popular DLP solutions available are Symantec DLP, McAfee DLP and Digital Guardian DLP. The following section explains how various DLP products in the DLP suite work and how they interact with each other by taking an example of the McAfee DLP solution.

4.7.1 Survey on McAfee DLP

McAfee Corp, a part of Intel Security Group, developed a DLP solution known as McAfee Data Loss Prevention, which protects sensitive data by identifying, monitoring, and preventing the data within a network. It helps in understanding the type of data and how the data being accessed and transmitted. Moreover, this McAfee DLP is a suite of five products. Each product has to protect different kinds of data in the organization. They are: [17]

1. **McAfee DLP Endpoint**: This DLP product focus on the prevention of data loss at endpoints. It inspects and controls the content of the documents and user actions on endpoint devices like workstations.
2. **McAfee Device Control**: It mainly focus on preventing data loss due to unauthorized use of removable media by blocking the device entirely or partially. It controls the use of any storage devices connected to the end devices.
3. **McAfee DLP Discover**: This module is responsible for the detection and protection of sensitive data. It identifies the data by scanning file repositories and then protect the data.
4. **McAfee DLP Prevent**: It mainly concentrates on the protection of web and email data. It prevents data loss by integrating with an MTA server or a web proxy to monitor email and web traffic.
5. **McAfee DLP Prevent for Mobile Email**: It is integrated with the Mobilelron Mobile Device Management servers to analyse the email traffic coming from the Microsoft Office 365 or Microsoft Exchange ActiveSync to generate and record the incidents in McAfee ePO for a subsequent case review. McAfee ePO software is a centralized management platform for prioritizing the alerts and ensuring that security tools work together with orchestrated controls.

McAfee DLP gathers entire data and categorizes it based on the vectors: Data-in-Motion, Data-at-Rest and Data-in-Use. The below table shows the suitable DLP product for various data vectors.

Data Vector	Description	Suitable Products
Data in Motion	Monitor live traffic on the network. Collected traffic data should be analysed, categorized and stored in the McAfee DLP database.	McAfee DLP Prevent McAfee DLP prevent for Mobile Email
Data at Rest	Monitor data present in the files and repositories. MacAfee DLP takes remedial actions after scanning and tracking data.	McAfee DLP Discover McAfee DLP Endpoint
Data in Use	Monitor user actions on endpoints. And control the removable devices by blocking or taking specific actions.	McAfee DLP Endpoint McAfee Device Control

Table 2 McAfee DLP Products [17]

How McAfee DLP products Interact

All McAfee DLP products should be installed to get a full feature of the DLP product suite. Each of these products needs to be installed and configured correctly in a suitable location in the organization.

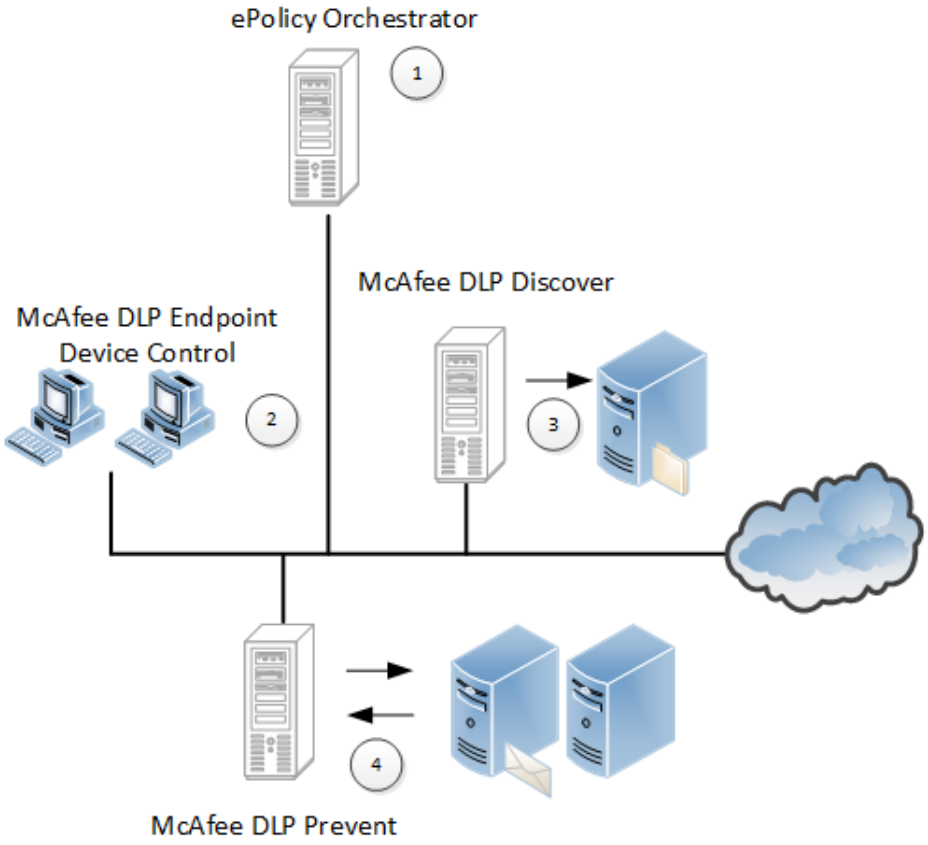

Figure 11 Interaction of McAfee DLP products [17]

As shown in Figure 11, ePolicy Orchestrator software provides a management platform to IT administrators for the configuration of policies and incident management for all its DLP products. This software also provides the automated management capabilities to identify, manage and respond to security threats by sending alerts and security responses based on the security incidents in the network.

The components such as McAfee DLP Endpoint, Device Control will perform their operations as discussed previously and interact with the ePolicy Orchestrator for incident reporting. McAfee DLP Discover monitors and scans all the documents available locally or in the cloud and reports to the ePolicy Orchestrator. [17]

The McAfee DLP Prevent analyse the messages received from the Message Transfer Agent servers (MTA is an application in the Internet Message Handling System for transferring the emails from sender's to receiver's computer) and combine these messages with the suitable headers based on the configured policy and send these emails to the single MTA server. In the case of web traffic, McAfee DLP Prevent collects the web traffic and analyse it to decide whether it should be allowed or blocked and send back the web traffic to the web proxy server.

Simultaneously, the McAfee DLP Prevent for mobile email receives emails from the MobileIron Sentry server (a secure gateway that provides conditional access to the apps and services on the premises) and analyse the attachments and contents of the email. Based on this analysis, incidents are created, or evidence is saved based on the protection rules. [17]

4.8 Best Practices for Data Leak Prevention

As discussed in Chapter 3, "Data Leak Threats," there are so many threat vectors resulting in data leaks. Although using a DLP solution in an enterprise gives data protection, there is a chance of data leakage. So, only utilizing a DLP solution cannot provide complete protection to the data. It is needed to take some extra measures to solve this issue, such as following security strategies and implementing the technologies like firewalls, Intrusion Detection Systems and Intrusion Prevention Systems.

For applying the best prevention measures to deter data leaks, a threat should be analysed before implementing those measures. For example, the data leak may be caused due to either insider threat or outsider threat. In another way, this leak can also be an intentional leak or an unintentional leak (also known as accidental leaks).

After detecting a threat, quick preventive actions are to be taken to stop the data leak. If it did not show any effect, the incident response team should try to decrease data loss by following the post-incident activities.

Some of the best practices for prevention of data leaks are discussed below.

4.8.1 Least Privilege Policy

The name Least Privilege itself shows that users should be granted only the rights and permissions required to perform their job duties. This principle is used to control access to the data. For example, an employee is working in the accounts department who has read access to all payroll files to perform tasks related to his/her job.

In this case, the management should give that employee only read access to the payroll data, required to perform the duties, but not with access to the data not related to his work, such as confidential information. Suppose the administrator incorrectly configured the accounts by ignoring the least privilege principle.

In that case, the user might have access to all the sensitive data such as trade secrets in the organization. Thus, the employee could steal the data and send it to outsiders to get financial benefits. Alternatively, the employee may leak the data accidentally into the public network. This leakage of sensitive data will make the company's secrets accessed by its competitors, which could finally impact its business.

This principle also ensures the prevention of access to sensitive data by providing only non-sensitive data when an attacker hacked the user account.

The Least Privilege policy can be implemented by creating different accounts for different purposes such as User accounts, Privileged accounts, Guest accounts and Service accounts. [11]

- End-user accounts are also known as Standard accounts. These accounts are created for the regular users performing normal duties that require access to normal data.
- A privileged account users have some elevated privileges than regular users. For example, an administrator in the Microsoft environment has a domain administrator account with special admin rights to control the systems in the domain.
- Guest accounts are created for the temporary users or users wanted to use the system without creating an account. These users have minimal rights and permissions. A contract employee from an agency can use this account for short and temporary access

to the organization's network. However, these accounts are commonly disabled because an outsider could access the network or systems using these accounts, leading to data exfiltration.

- Service accounts are created to run some applications and services. These are just regular end-user accounts created for service such as SQL server, which needs to access resources on the server by using this account.

Based on these above types of accounts, each user group is assigned with required privileges, and then users will be allotted to these groups based on their job responsibilities. Hence, the Least Privilege principle reduces the security risks associated by reducing access to a limited amount of data.

4.8.2 Data Encryption

Data encryption is one of the best practices to prevent the loss of data confidentiality. Both data-in-rest and data-in-motion can be protected using encryption techniques.

Encryption converts plain text data into ciphertext by using a key (a unique and unpredictable random string of bits used for encrypting and decrypting the data). The ciphertext is then converted into plain text using the same key when the user wants to read it. Later, this ciphertext should be decrypted using the same key to obtain the data in plain text.

There are several encryption algorithms such as Advanced Encryption Standard (AES), Triple Data Encryption Standard (Triple DES) and Rivest Shamir Adleman (RSA).

Encryption ensures that the sensitive data is not accessible by unauthorized persons even if the attacker gains access to the storage device. The Microsoft NTFS contains Encrypting File System (EFS) technology available on most windows systems. The EFS encrypts the files and documents stored in the system. When an authorized user opens the encrypted file, the EFS decrypts the file in the background and gives the file's unencrypted copy to the user. If the user performed any modifications to the file, EFS saves the changes to the encrypted data parallelly. [18]

BitLocker is another encryption technology from Microsoft which is used to encrypt the data on windows-based systems. The BitLocker provides the full disk encryption, whereas the EFS provides individual file and folder encryption rather than the entire drive. Moreover, both EFS and BitLocker could be used simultaneously based on the requirement. [18]

Linux-based systems also support file and folder level encryption. GNU Privacy Guard (GPG) is a command-line tool for encryption of data in a Linux system.

Data in the databases can be protected with encryption. Many database applications such as Microsoft SQL Server and Oracle Database can encrypt the entire database. However, it is not required to encrypt all the data elements in the database. [11]

For instance, there is a database containing a customer's table. This table may have multiple columns such as Customer identification number, name, credit card number, expiry date and card verification value (CVV). In this case, sensitive data columns like credit card number, expiry date and CVV in each record are encrypted instead of encrypting the entire database. Moreover, the full database encryption required more processing power.

4.8.3 Clear sensitive data from Non-Critical Systems

Data leaks can be prevented by not placing any sensitive data on non-critical systems such as workstations, laptops and removable media devices. If this data is placed on these devices, there is a high possibility of data theft when an attacker performed a random malware attack on the organization.

Usually, the non-critical systems are less protected when compared with the critical systems like servers. For example, an employee is working with a copy of PII in his system. If the system got attacked by the malware, the PII data could be exfiltrated, leading to a data leak.

The systems can be cleaned by removing sensitive data from non-critical systems to lower the risk of a data leak. This can be achieved by placing confidential data on well-protected systems.

Every organization needs to implement a 'Clean Desk Policy', to make sure their employees organize their desks properly. This policy ensures that no sensitive data is placed on the working desks and prevents the unintentional disclosure of data and data thefts.

Consider the situation in which an officer placed some papers containing confidential data on his desk. Anyone who enters the cabin to meet the officer could see the documents placed on the desk. Moreover, they could take pictures of those documents or take some of the documents with them if the officer went out. The office desks should be clean and well maintained by placing documents in a secure place. [11]

4.8.4 Malware Protection

Malware attacks are one of the outsider threats causing the data leaks. Safeguarding the endpoint systems and the network of an organization from malware infections is one of the best practices for preventing data leaks.

Quick Identification of malware attacks is vital to stop the spread of malware over the network. As the identification time of an attack increases, the time required for an attacker to breach the security systems will decreases.

Usually, DLP solutions cannot protect the systems and network from malware attacks. However, the Intrusion Detection System (IDS) monitors the network's activity and alerts if it detects a threat. Some IDS can block suspicious traffic and malicious activities. Moreover, Intrusion Prevention System (IPS) monitors the network activity and raises an alert if it detects a threat and prevents the threat from causing damage to the network.

Implementation of firewalls is helpful to prevent malware from entering the systems. Firewall filters suspicious traffic coming into the network or going outside to the network based on the predefined rules. As mentioned in chapter 3, "Data Leak Threats", a malicious insider can upload sensitive data to the FTP server using an FTP client in the internal network. Creating a rule on the firewall to block FTP traffic going outside from the internal network and also traffic entering into the network to prevent the malware.

The malware infection caused due to removable devices can be prevented by using an anti-malware solution on the system. Consider a scenario where an attacker intentionally drops a USB flash drive infected with a worm (a type of malware) on the company premises. An employee might notice the flash drive and curiously insert the flash drive in his workstation. The worm gets installed automatically within seconds on the user system. It then replicates itself and infects the remaining systems in that network. The worm can cause ransomware attacks or even exfiltrate the data. An anti-malware solution should be installed on the system to detect and quarantine the worm.

Unified Threat Management (UTM) can be implemented in an organization instead of implementing IDS, IPS, firewalls and anti-malware separately. UTM is a single security solution that contains multiple security functions such as Uniform Resource Locator (URL) filtering, content inspection and malware inspection. It acts like a proxy server and block access

to the sites based on the URL filters. UTM inspect the data stream coming into the network for any malware and malicious content and block them. Furthermore, it detects data breaches and prevents the data from exfiltration. Basically, it provides all the functions provided by DLP, IDS, IPS, firewall, anti-malware, content filtering and data leak prevention. [11]

Some of the popular UTMs are Fortinet UTM, WatchGuard's Firebox UTM and Sophos UTM.

4.8.5 User Training

As we know, most of the data leaks in any organization are due to the employee's mistakes. A user who lacks basic cybersecurity knowledge might click on the malicious link in the phishing email, which could infect the system with malware or redirect him to a phishing website. The user is requested to enter his credentials to log in. Attackers then collect those credentials and use them to do illegal activities. So, proper awareness training on the cybersecurity threats and security policies should be given to the employees to avoid making mistakes that could lead to a data leak.

Role-based awareness training is followed in many companies to give training to the individuals based on their roles. The following roles are used to provide role-based training: [11]

- **Data Owner** needs to understand their responsibilities related to their data, such as classification and labelling of data. They also ensure the implementation of proper security practices to protect data.
- A **system administrator** is responsible for system's security and requires training to understand the capabilities and vulnerabilities and make sure that the system is running in a safe state.
- **System owners** are responsible for ensuring that system admins have the skills and knowledge to maintain a system. Moreover, they are a high-level executive in a department.
- **Users** who are regular end-users should have basic knowledge of malware threats and phishing attacks. Training is given on various topics and delivered online, on websites and in a classroom.
- **Privileged user** has more rights and permissions than the regular end-user such as administrators. They need training on data classification and labelling.
- The **executive user** needs high-level knowledge of the risks that the organization might face and is responsible for the overall security awareness training program.

- The **incident response team** needs to be trained rigorously to perform incident response activities.

By providing awareness training based on the job roles, most accidental data leaks can be avoidable.

4.8.6 System Hardening

Sensitive data could reside temporarily on several systems during transmission or even in usage. All the systems, including external systems accessing data through a remote connection, should be secured based on the data that the system could access.

Hardening is the process of securing a system or an application by changing its default configuration. Network devices, applications and operating systems (OS) can have default usernames and passwords. Generally, most of the newly installed routers have some basic default settings such as admin login username as ‘admin’ and password as ‘password’.

If these credentials are not changed, an attacker could easily guess these credentials and log in to the router to know the Wi-Fi password. The attacker then uses a packet sniffer like Wireshark to sniff the data packets to capture the user’s data on the local network. So, the router default configuration should be changed as soon as installing the router.

Hardening secures the system from vulnerabilities and misconfigurations by updating and patching the system regularly using a patch management system. Every system in the organization should be keep updated with new versions. Moreover, tools such as anti-malware should be updated automatically with new signatures.

Some of the technique used to secure the OS and applications are mentioned below: [19]

- **Hardening:** Hardening of a system involves removing default configurations, performing vulnerability assessments and performing updates to the system
- **Least Functionality:** Configuring servers only to provide essential applications and services.
- **Application Whitelist:** Only the applications in the list are allowed to run by the OS.
- **Application Blacklist:** Only the applications mentioned in the list are not allowed to run, but the remaining applications are permitted to run.
- **Disabling Unnecessary Services and Ports:** The services and ports which are not necessary should be disabled.

- **Trusted OS:** OS should meet all the requirements set by the authorities (Windows 7 or higher version, Mac OS X 10.6 or higher version).
- **Two-factor Authentication:** Two factors such as ‘something you know’ and ‘something you have’ are used for authentication. An access card, along with the password, comes under two-factor authentication.

4.8.7 Destruction of Data

Every organization generates lots of e-wastage containing devices like computers, servers and hard drives. These devices may contain some sensitive information. So, all the data should be deleted before disposal of the hard drives. Simply deleting the files or formatting the drive does not remove the data permanently. The deleted data from the drives can be recovered using several data recovery applications.

So, the data destruction techniques can be used to destruct the data permanently and prevent it from recovery. These techniques are meant for the hard drives and the data in paper format. If these papers are disposed of in the trash without following any measures, dumpster divers can search through the trash and gain this valuable information.

So, the following data destruction techniques can be used to destroy and sanitize media: [11]

- **Purging:** It is a general term to indicate that all the valuable data is removed from a device.
- **File Shredding:** Removing all the file remnants by repeatedly overwriting the file’s space with 0s and 1s.
- **Wiping:** Completely removing all the disk remnants using a disk wiping tool by writing 0s and 1s repeatedly multiple times on the disk.
- **Erasing and Overwriting:** Removing data from the solid-state drive is not possible through wiping because it uses flash memory. So, the sanitization process is used in which drives are destroyed physically.
- **Paper Shredding:** it is the most common type of paper destruction method. In this method, papers are passed through cross-cut shredder machines to cut into fine pieces.
- **Pulping:** The shredded paper is converted into puree or pulp.
- **Degaussing:** The process of passing hard disks through a degaussing field produced by a powerful electronic magnet. It makes the data on the disk unreadable.
- **Burning:** Papers are destroyed by burning in the incinerator.

4.8.8 Perform Penetration Tests

Penetration testing should be performed frequently in every organization to know the security posture of a system or a network. Penetration Testing or Pen Testing is attacking a system or a network to detect any vulnerabilities and exploit those vulnerabilities to check the organization's security posture.

The pen test involves intrusion attempts to check security measures such as IDS and IPS to know any potential weaknesses. A pen tester can perform this within strictly defined boundaries because it can disturb the actual operations of the system. In some cases, tests are performed in the simulations or on the test systems to prevent causing damage to real systems.

A pen test is conducted in several stages. They are: [11]

- **Passive Reconnaissance:** The process of collecting the information about the target systems passively by using open source intelligence such as websites, social media and news reports and performing IP look ups and Whois lookup to know the IP address and domain names holder's information.
- **Active Reconnaissance:** Gathering information about the target using network scanning tools and vulnerability scanners such as Nmap and Nessus tools.
- **Initial exploitation:** Exploiting the vulnerabilities detected through passive and active reconnaissance and allowing the testers to gain access to the system. Metasploit is the penetration testing framework that is used to exploit the known vulnerabilities.
- **Privilege Escalation:** Gaining more privileged access to the system after exploiting the system using privilege escalation techniques like bypassing user access controls. That is gaining administrative privileges after successfully gaining regular user access.
- **Pivot:** Process of gathering more information about all other systems using various tools after gaining access to a system.
- **Persistence:** Maintaining access and staying on the network for several weeks or months without being detected. This can be done by creating backdoors into the network by modifying services to connect the system.

4.8.9 Defense in Depth Strategy

The defense-in-depth strategy is implementing several layers of protection to defend an organization's network from outside threats. A few protection mechanisms such as firewall, anti-malware, and IPS will not be sufficient for overall security. So, several layers of security

measures should be followed at every phase. If one layer fails, then the remaining layers would protect the network.

Defense in depth uses control diversity security measures such as administrative controls (policies, guidelines, risk assessment, vulnerability assessments and training), technical controls (Encryption, DLP, IDS, IPS, firewalls and access controls) and physical controls (security cameras, guards, doors, locks and fence). These controls are implemented at various locations to protect different systems and the data, the most valuable asset of any company.

There are several types of defense-in-depth models such as 4-layer, 5-layer, 7-layer and 13-layer models. Any model can be selected based on the company's requirements. [20]

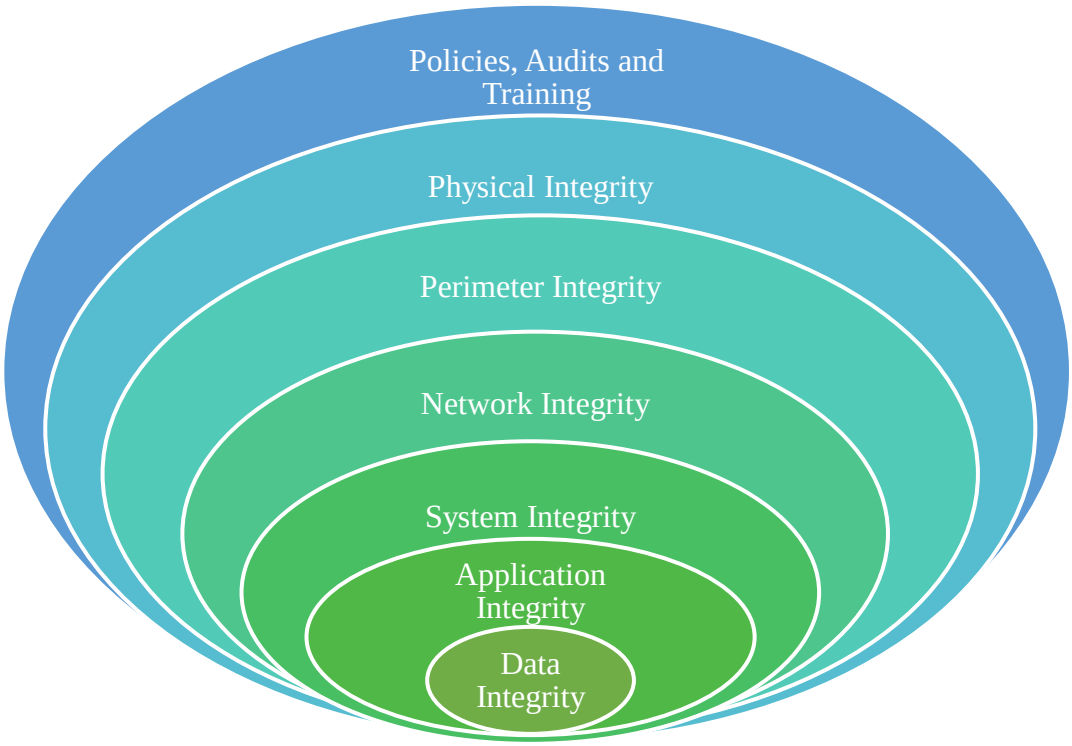

Figure 12 Model of Defence-in-Depth Strategy [20]

Figure 12 shows the 7-layers model of defense-in-depth strategy. Each layer of protection is explained below: [20]

1. The policies help understand the business structure and define and implement rules and security baseline in an organization. A strong data recovery plan will minimize the

damage caused due to data leaks. Audits should be conducted frequently to check whether the policies are enforcing correctly or not. Training should be given about the policies to every employee.

2. Physical Integrity prevents unauthorized persons from accessing the organization's assets, such as IT resources and physical property. Physical security measures such as guards, walls, doors, locks, and cabinets are used to protect the resources.
3. Perimeter Integrity ensures the implementation of security measures on the network devices such as routers to protect from attacks and intrusions. Firewalls, IDS and IPS are helpful to achieve perimeter integrity.
4. Network Integrity protects the internal network by monitoring and managing the traffic within the network. Network devices like switches are included with VPNs, VLANs, MAC filtering and disabling unused ports.
5. System Integrity protects the end systems and ensures that security policies are implemented on all the end devices such as servers, workstations and laptops. And can be obtained by hardening the hardware and operating system.
6. Application Integrity protects the software application in a system by configuring proper security settings. Technical controls such as authentication, access controls and encryption are used to secure general, email and database applications.
7. Data Integrity deals with data protection by classification and validation of data in the database applications and place in a secure location in an encrypted format and only accessed after authentication.

If an attacker tries to access the database's data, he needs to bypass all these security layers before accessing it, which is very difficult. Even though the attacker bypassed any layer, there are still several layers of protection between the attacker and the data. Therefore Defense-in-depth is the best strategy to follow to secure the data.

4.9 Incident Response Plan

An Incident Response Plan is a roadmap to follow when a data breach occurs. This plan can save much time and make data recovery easy. However, it should be created in advance to use this in a crisis. The incident response plan contains the following tasks:

4.9.1 Defining a Data Breach

It is required to determine what type of breaches might activate the response plan. For example, a company might experience phishing email events that may have a negligible effect on their

operations. In comparison, a data breach could disrupt the company's operations, which needed to follow a response plan. This plan identifies these incidents and makes them into groups using category definitions. [21]

4.9.2 Response Team

A response team is required to implement the response plan. It should include at least one representative from each department who can take responsibility seriously and have overall knowledge about its operations. Each member is assigned a role based on their department.

For example, a data security employee should determine how the data leak occurred, and a risk management employee needs to notify the insurance company which provided a cyber liability policy. The team's size varies from each organization and depends on the size and complexity of the business. [21] Furthermore, this team often has extensive training on detecting and validating an incident and collecting and protecting evidence.

The following departments should be involved in the response team: [21]

- **IT security:** Detect and respond to the breach.
- **Legal and compliance:** Determines and maintains data retention policies (a policy that defines how to save and dispose of data based on the regulatory standards) and informs the appropriate authorities.
- **Public relations and marketing:** Deals with customer identification and communications tasks.
- **Sales:** It leads with relationship management.
- **Executive:** Coordinates high-level response efforts.

When a mega data breach occurs, the regular response team is not enough to deal with a data leak. Outside consultants like law enforcement officers, recovery experts and attorneys are required. [21]

4.9.3 Action steps

After identifying a data breach, the response team follows step-by-step instructions to escalate it. Based on the severity of the data breach, security personal in the team should notify about the potential effect on the critical operations to their executives. The team members should document all the actions performed during the incident response.

This ensures that the team members followed all the instructions in the plan, and it is valuable in the post-breach evaluation. The documentation also helps report the data leaks to the authorities if it is involved in the data protected by law, such as PII. [11] [21]

4.9.4 Exercises

Exercises should be conducted regularly to prepare for incident response. After successfully containing the data breach, the team should conduct a debriefing session asking the team members to run steps and the lessons learned from the process. They should discuss the problems raised during the process to adjust the plan as needed. [21]

4.10 Response to a Data Breach

Every organization should follow an incident response process to handle data leaks effectively. This involves multiple phases and starts with creating an incident response policy and incident response plan. This response plan should be created ahead of the actual incident response process. By using this plan, employees are trained with the necessary tools for handling incidents.

The response process to handle the data leaks has the following phases:

4.10.1 Preparation

This phase starts before an actual incident happens. It establishes and maintains a data breach response plan to prepare for a cyber emergency. The response teams are created with personals from every department, and the IT resources are allocated to the most crucial departments. Exercises and awareness activities such as security awareness training sessions and simulated cybersecurity incidents are conducted once or twice a year. [22] Prevention is the primary goal of this preparation phase.

4.10.2 Identification

In this phase, the response team identifies the type of data leak and its severity. This phase also determines which department to involve in the response activities. All events raised in the organization are not security incidents. [11]

A security incident is reported when the team verifies the event. [11] For instance, a DLP may raise false alarms of a breach, but they need to investigate whether it is an actual incident or not. When the team identifies an incident, they need to isolate the affected system to protect the remaining systems.

After detecting a data leak, the company should notify their customers and the authorities based on the leakage data. If the data is legally protected information such as PII and health records, the company must inform its customers and authorities. Some sectors like retail and Payment card industry (PCI) compliance must inform their customers. Simultaneously, other sectors like government and healthcare sectors should inform their customers and their regulatory authorities. [22]

Suppose the incident involves losing a company's data, such as trade secrets and intellectual property. In that case, they no need to announce to the public because the damage is only to that company. [22]

4.10.3 Containment

The response team tries to mitigate the data breach's impact by isolating or containing it. The team's goal is to stop the spread of the breach and secure all the data. After isolating all the infected systems, any password or encryption keys should be changed immediately to stop unauthorized access by the attacker.

If the data breach occurred due to a malware attack, the infected system must be isolated as quickly as possible. For example, the malware-infected system is isolated physically by merely unplugging its network card from the network. Moreover, it can also be isolated by changing the access control lists on the routers or firewalls.

Then the required resources are allocated to remove the malware from the system. An effective data backup strategy helps to provide the best procedures to retrieve the data. If it is ransomware, the best response is not to pay to release the data. Instead, the most recent backup data could be used to restore the operational state. [11] [22]

4.10.4 Investigation

In this phase, the team starts its investigation to know the cause of the data leak. All this investigation process, actions were taken by the team is documented for future use. The law enforcement is then involved in the investigation, and the documentation report is submitted to the authorities. Furthermore, if any additional support is needed for the investigation, the executive leadership teams and legal counsel are consulted. [22]

Finally, the collected digital evidence of the breach is protected by following a set of instructions and approved methods such as order of volatility and capturing system image. The

order of volatility is collecting evidence starting from the most volatile memory to the least volatile memory. [11] For example, Random access memory (RAM) is the most volatile, containing valuable evidence. So, it must be captured before turning off the system.

The collected evidence should not be modified. The chain of custody assures that the evidence has been appropriately handled without any changes after the collection.

4.10.5 Recovery

In this phase, the team concentrate on recovering all the system to the normal state and verify them to check whether they are operating normally. The recovery contains the operations such as rebuilding the affected systems from the images and restoring the backups. Moreover, the vulnerabilities that caused the data breach should be removed by updating the system to prevent them from happening again in the future. [11]

4.10.6 Lessons Learned

After successfully handling a data breach, the incident response team conduct a lessons learned review. Every incident gives some valuable lessons to the organization. The team creates a lessons learned report by including content such as the new challenges the security team experienced and how they solved those challenges. [11]

This report helps in preventing the reoccurrence of the data leak due to the previous threat. This review can provide training to the new team and update the incident response policy. [11]

Chapter 5: Conclusion

Every organization, whether it is a small or big one, has been depending on the data for every operation they perform, so data in any form is a crucial asset. The leakage or loss of data permanently could cause a devastating effect on companies, governments and even individuals.

As the data volume is increasing rapidly, data leaks have been happening continuously since 2005, when the first data breach had occurred. Furthermore, the occurrence of data leaks has been increasing year by year.

Although most companies are following data protection mechanisms, data exposure continued to occur due to many factors, which may be internal or external threats. Based on the analysis, most of the data leaks are internal leaks due to poor security practices rather than external threats.

As we know, the data protection mechanisms are proliferating; simultaneously, the threat vectors are also developing continuously. So, there is necessary to use more advanced data leak prevention techniques to protect the data.

Although a typical DLP solution takes preventive measures, it cannot provide complete protection from data leaks. Every organization handles different types of data and requires suitable prevention measures based on its organizational structure. So, the company's network should be build based on the Defense-in-depth strategy. This layered approach will prevent most of the threats and safeguard the organization's network.

Despite the implementation of prevention mechanisms, there is still a chance of data leakage. An incident response team should be included in every organization to conduct required activities to stop the data leak and recover it.

The security threats causing the data leaks may evolve in the future, and it could become more difficult to mitigate these threats. So, it is needed to adopt new data protection approaches. Finally, the prevention of data leaks is not only the security team's responsibility; every individual must understand the basics of cybersecurity to avoid making mistakes and responsible for data protection.

Part II

Chapter 1: Introduction of Data Harvesting

1.1 Data Harvesting

Data Harvesting, also known as web scraping, is a process of collecting a large amount of unstructured data such as text, photos, and email addresses automatically from several websites using malicious bots (scripts) to use that data for other purposes. It mainly focuses on the online data placed on the websites publicly. [23]

In other terms, data harvesting can be defined as the process of representing, analysing and extracting actionable patterns and trends from raw social media data. It requires human data analysts and automated tools to go through the massive social media data to discover patterns and trends. The social media data includes connections, usage, online behaviour and online buying behaviour of the users. [24]

1.2 History and Evolution of Data Harvesting

Data harvesting incidents have been happening for several years, along with data leaks. In 1993, Matthew Gray created the first non-malicious bot known as World Wide Web Wanderer. This bot was released to traverse the web and discover new sites, but it did not reach every website. [25]

However, the first malicious bot is an e-commerce web scraping bot known as Bidder's Edge, revealed in the early 2000s. This bot was created to aggregate competitor's pricing in the auction sites. [26]

Later several web scraper bots are created to harvest the data from the sites are increasing year by year. Some of the data harvesting case studies are described in the following sections:

1.2.1 Facebook Cambridge Analytica Scandal

A personality quiz for Facebook was developed by a man named Aleksandr Kogan in 2013. Based on the Guardian report, this app was installed by 300,000 people giving Kogan access to a million people's data. [27]

Later, Kogan alleged that he shared that data with Cambridge Analytica, a data mining and data analysis company. That company used the data from Facebook to target US presidential election voters in 2016. [27]

Facebook responded to the Cambridge Analytica scandal after several days. Mark Zuckerberg released the timeline of events and steps Facebook would follow to prevent this from happening again. [27]

Facebook's solution involved three steps. First, they were conducting a "full audit of any app with suspicious activity", second they would "restrict app developers access to data," and finally they would help users "to learn which apps users have allowed to access their data". [27]

1.2.2 US Voters Data Exposed

In 2017, the Republican National Committee hired a Deep Root Analytics company, which compromised around 198 million American Voter's personal information. They gathered that data to influence the voters, working with two other companies known as Targetpoint Consulting and DataTrust. [28]

UpGuard, a cybersecurity research company's researcher Chris Vickery found 25 Terabytes of data was exposed on an unsecured Amazon Web Services S3 bucket, making 1.1 Terabytes of data available for download. The leaked data contained personal information such as names, contact numbers, DOB, addresses, registered party, and details related to ethnicity and religion. Later, Deep Root Analytics was charged in a lawsuit. [28]

1.2.3 South Africa Data Leak

A database backup named masterdeeds.sql containing personal information and 60 million unique ID numbers was found on an unsecured public server in South Africa. The data records in that server exceeded the entire country's population, and it contained information about dead people. Moreover, it included the ID numbers of 12 million minors of the country. Criminals could use this data to commit fraud or steal the identities of the people. [29]

Cybersecurity expert Troy Hunt, founder of "HaveIBeenPWned.com", warned about that data exposed over seven months. This database was created by a company named Dracore, but one of its clients, Jigsaw Holdings, disclosed the data. [29]

1.2.4 CRM Email Spamming

Although Data aggregation companies are legal, some of these companies illegally harvest data. The case of City River Media (CRM) is an example of a massive illegal spamming

operation. CRM exposed the data of 1.4 Billion user's email accounts along with names and IP addresses. [29]

Mackeeper (Security Research center), CSOOnline (security and risk management company) and Spamhaus (threat intelligence firm) were investigated this leak and found that misconfigured Rsync backups making the data vulnerable. [29]

CRM was considered to be a legitimate marketing company, but was later found to be a spamming company, sending billions of spams emails every day. Chris Vickery, a MacKeeper's researcher, accessed CRM's logs, accounting details, domain registration records, scripts, business affiliations and production notes. Later he submitted these details to the authorities. [29]

1.2.5 Alteryx Data Leak

In 2017, UpGuard found that a data analytics company known as Alteryx caused the exposure of 123 million American household's data by placing it on an unsecured data repository. The data repository included addresses, mobile numbers, mortgage details, financial information and purchase history. This data could have been accessed by anyone who had an Amazon Web Services account. [29]

Fortunately, this data was available only for a few days, but still long enough for criminals to download the sensitive data within a short time span. [29]

Chapter 2: Analysis of Data Harvesting

2.1 How Data Harvesting Happens?

As mentioned in the previous chapter, “Introduction of Data Harvesting”, Data harvesting, also known as web scraping, uses automated processes to gather massive information from websites and social media platforms to publish it elsewhere. It is an illegal activity performed without the consent of the site’s owners.

Commonly, websites contain a large amount of valuable data. However, this data and websites are meant for the human end users, and it is difficult to scrape the website’s data, so specialized tools are created to perform web scraping.

Web scraping can be performed in many ways, such as using online services, particularly Application Programming Interface (API), and even creating a web scraper from scratch. Most of the social media platforms have APIs that permit access to their structured data. However, some websites do not allow their users to access data in a structured format due to the lack of advanced technology to access the data. In this case, web scraping is the best way to collect the website’s data. [30]

Web scraping can also be performed by listening to the server’s data feeds during the communication between the server and the client. For instance, JavaScript Object Notation (JSON) is the lightweight data-interchange format and used as the transport storage mechanism. So JSON format is used for scraping in between web server and client. This technique can be used when the website does not offer an API. The process of web scraping requires the following elements: [30]

2.1.1 Web Crawler

A web crawler (crawler or spider) is an artificial intelligence algorithm that browses the World Wide Web (WWW) to search for required data by following the links across the internet. Search engines and other websites use these crawlers to update their web content or indices of the other site’s web content. They crawl all the pages one at a time in the website until all the pages have been indexed. Moreover, they consume the resources on the visited sites without approval.

2.1.2 Web Scraper

A web scraper is a tool that deploys bots to extract a specific type of data from the website or an API. These scrapers can be used for a good intention or a bad intention. Web scrapers are used for creating visibility to the website or content. However, the malicious scraper will send

bots to steal original content from the websites. The advantages and disadvantages of web scrapers are described later in this chapter.

2.1.3 Bot

A bot is an application that runs automated tasks over the internet. There are two types of bots, known as Good bots and Bad bots. Many studies show that over 50% of all internet traffic is of bots. However, they focused on the malicious bots which need to be detected and blocked before causing any harm.

- **Good Bots:** The bots that are beneficial to the company's business or even for the individuals are Good Bots. For instance, a search engine like google uses GoogleBot to crawl over the internet to search for a website based on the phrase entered in the search bar and shows the result in the results pages. The good bots work based on the webmaster's rules to control the activity and indexing rate. These rules are defined on the website 'robots.txt' file. These bots can also be prevented if they are not applicable to the business. There are several types of good bots, such as partner bots, social network bots, website monitoring bots and aggregator bots. [31]
- **Bad Bots:** Bots that are used to perform malicious activities are known as Bad Bots. Cybercriminals mainly use bad bots to steal the content of the website. Moreover, they may affect the servers by using more resources and bandwidth of the servers. [31]

In general, data harvesting involves mainly three tasks. They are explained below: [32]

- **Retrieving Data:** This is the process of finding the webpages containing valuable information and storing it locally. This process required tools such as a web crawler for searching and navigating through the web.
- **Extracting Data:** Identifying the valuable data from the retrieved webpages and extracting it from the pages into a structured format. The extracted data can be analysed further by using the tools like content spotters, adaptive wrappers and parsers.
- **Integrating Data:** This involves filtering, transforming, combining and refining the extracted data from several webpages and structuring the refined data based on the desired output. It is essential to organize the extracted data to allow data mining tasks for further analysis.

Data harvesting's primary purpose is to combine as much information as possible from the websites from many domains to create a massive structured knowledge base that allows querying for information similar to the regular database.

More specifically, it can be performed using a web scraper that extracts all the data or the website's specific data. However, it is best practice to specify the data during web scraping. The web scraping is done by following the below steps: [30]

- Web scrapers is provided with one or more URLs to load the web sites.
- Web scraper then loads all the HyperText Markup Language (HTML) code into it. The data in the CSS and JavaScript elements can be extracted by using advanced scrapers.
- Web Scraper then extracts all the data or specific selected data from the HTML code.
- Finally, the extracted data is given as output in any format such as Comma-Separated Value (CSV), excel spreadsheet and JSON file.

2.2 Web Scraping Techniques

Web scraper can be performed by using several techniques such as simple copy and paste, HTML parsing. Some web scrapers use artificial intelligence for scraping the data. Some of the techniques are explained below:

- **Copy and Paste:** This technique is a manual extraction technique, in which the data is extracted manually by copying the website's data and pasting it into spreadsheets or text files. This is a simple technique, but it is effective when the websites prevent web scrapers. To check whether websites allow scraping or not by looking at 'robots.txt' file of the website. For example, the Bing website does not allow a path like "/hotel/reviews?", in this situation, the data can be scraped manually by copying and pasting into a text file. [30]
- **HTML Parsing:** It is an automated extraction technique. In this technique, the HTML pages are targeted. The HTML code is taken, and required information is extracted based on the user requirement. Parsing makes the data understandable for analysing part by part. This parsing of HTML pages can be done using XQuery and the HyperText Query Language (HTQL). [33]
- **DOM Parsing:** The Document Object Model (DOM) is used in XML files to define the style, structure and content of the file. Scrapers use the DOM parser to get an in-depth view of the webpage structure and get nodes containing information. A web browser can be embedded with this parser to get the entire page or even a part of the web page. [34]
- **HTTP Programming:** The socket programming is used to extract static and dynamic data by posting the HTTP requests to the webserver. This can be done by using

JavaScript and target linear or nested HTML pages. This method can also be used for text extraction, link extraction and screen capturing. [34]

- **XPath:** XML Path (XPath is a query language that works on XML documents) is used with DOM parsing to extract the web page and publish it on the destination website. XML documents are the tree-like structure, so the XPath language is used to navigate on the tree by selecting nodes based on the parameters. Moreover, finally extract the data from those documents. [34]
- **Google Sheets:** Google sheets are used as scraping tools. The function IMPORTXML (,) in the google sheets can scrape the web site's content when the scraper wants a specific pattern to extract from the website. [34]
- **Text Pattern Matching:** Website data is extracted using the 'grep' command in UNIX systems. The regular expression matching facilities of the programming languages like Perl and Python. [35]
- **Computer Vision Web-page Analysis:** Machine learning and computer vision technologies are used to identify and extract the data by interpreting the web pages visually as a human. [35]

Scrapers can use any one of the above techniques to scrape the data. They might no need to know all these techniques unless they want to scrape automatically by themselves. Moreover, they also use automated headless browsers (a browser that provides automated control of a web page through a command-line interface or network communication) such as Phantom.js and Slimmer.js for scraping websites.

2.3 Types of Scraping

Basically, the scraping can be done to extract different kinds of data such as websites contents, prices and advertisements. Based on the data, the scraping is classified into four types. They are explained below:

- **Content Scraping:** A standard website's content is scraped in this scraping. Media websites that post news, blogs, educational and research content are the most targeted for content scraping. This scraping will impact the company's revenue in many ways, such as copying the content and republishing on fake websites. Moreover, negatively impact the company's property and affect subscriptions and advertisement revenue. [36] For example, an online streaming website has massive content in the form of movies and TV shows. Suppose this content is scraped and placed on the torrent sites

for free. The streaming company will lose its subscribers and simultaneously lose revenue.

- **Price Scraping:** Scraper bots are used to monitor the competitive price illegally and track the information related to the price from the e-commerce, travel, and retail websites. Although this price data is publicly available, the scraper bots are used to undercut the competitor's price and grow their business. [36]
- **Advertisement Scraping:** Business competitors and fraudsters use scraper bots to scrape the advertisement listing from the websites. This scraping causes the site owners to be denied from realizing commissions and listing fees for sales using their websites. The ad scraping can also affect the credibility and the efficacy of the ad listing site, resulting in reduced user volumes. [36]
- **Schedule and Availability Scraping:** Hotels, car rentals, airlines and cruise lines use the Global Distribution Systems (GDS) to list the pricing data, availability of accommodation and travel packages. The travel agents use this GDS system to provide the best deals to their customers. This scraping allows the fraudulent operators to access legitimate agent's data to avoid the GDS query fee, which will lead to an increase in the higher GDS fee to the genuine operators. Furthermore, this scraping can also affect the bookings and SEO when the content is duplicated. [36]

2.4 Uses of Web Scraping

Web scraping is used to extract, review and sort massive amount of data quickly and systematically. And has many advantages in various fields, both professionally and personally. Some of them are explained below: [33]

- **Brand Monitoring:** A company can perform web scraping to collect the feedback of a specific service or a product, which helps find how customers feel about the product. For example, when the publisher wants to know the reviews of their book. They can perform web scraping to extract all the book reviews from several e-commerce websites such as Amazon and eBay. A company may also use web scraping to know about their competitor's products and services. [33]
- **Social Media Analysis:** Web scraping is used to know the customer trends and behaviour in the social media sites and know how they react to the campaign. In the Facebook Cambridge Analytica Scandal case study, the Facebook user's data is collected and used to influence the US presidential election's voters in 2016. [33]

- **Machine Learning:** Machine learning is the study of artificial intelligence that allows a machine to learn and improve its accuracy through experience rather than programming. A machine needs lots of data to learn, which can be extracted through web scraping tools. [33]
- **Financial Data Analysis:** Web scraping is used to know economic trends and maintain the stock market record in a usable format that could be used in the future. For example, an investing company needs an in-depth piece of data to assess before investing in a particular stock. [33]
- **Data Comparison:** Scraping can be performed on the APIs to retrieve particular data to compare each other. For example, most hotel price comparison websites such as 'trivago.com' use bots to gather hotel prices based on the required specifications.
- **Email Marketing:** Many companies depend upon web scraping techniques to collect email IDs from various websites and social media platforms. Then marketing and promotional emails are sent to these email IDs. [30]

2.5 Negative Effects of Web Scraping

Generally, web scraping tools are developed to extract the data for good purposes. However, unfortunately, these tools are used for data harvesting, which has always been a concern for website operators and data publishers. Web scrapers are used to do illegal activities such as stealing data from websites. If keeping the data loss aside, there are several adverse effects on the organization's business in many ways. Some of them are given below:

- **Poor SEO Ranking:** After scraping the website's content, this data will be posted on the other websites. This reproducing of data could affect the performance and Search Engine Optimization (SEO) ranking. The ranking given to a page is based on the position of web pages displayed in the search engine result. For example, if data is scraped from a website and placed on the other page, the SEO ranking of the reproduced page will get the first rank that means it will be displayed before the page containing the original content. [23]
- **Lost Market Advantages:** Competitors of a company might perform the data harvesting to scrape the valuable data from the website to gather intelligence of the business which could result in the loss of business. [23]
- **Decreased Website Speed:** When the web scraping is performed on the website continuously, its performance would be decreased, which could affect the user

experience. Due to the increased web scraping traffic, the company's network bandwidth cost will raise. [23]

- **Fake Websites:** The original website's data is harvested and used to create similar fake websites. For example, an e-commerce website's HTML source code is extracted using the web scrapers and this code is used to replicate the site. Moreover, this website is launched with a similar domain name, which is known as Typosquatting. [37] When the shoppers typed the website's name in the search engine, there will be high chances of showing the fake website in the search results. Unsuspecting customers may use this fake website and enter their valuable information on this website. Later this data could be used to commit fraud.
- **Intellectual Property Issues:** The malicious bots used in the data harvesting could steal confidential data such as trade secrets, product information and even pricing. The information is then used to perform a price-cutting strategy. The price of the product is decreased based on the competitor's product pricing. This strategy makes the values and ranking of the product in the market is thrown off. [37]
- **Analytic Issues:** Hundreds of bots across the internet may try to scrape a website's data at a time. This interaction of the bots could affect the tracking ability of the website. The website cannot be able to detect which traffic is from real users and which is from bots. This flawed data could lead to making wrong marketing decisions. [37]

2.6 Web Scraping Tools

Web Scraping tools are also known as data extraction tools, or web harvesting tools developed to extract specific data from the websites. These scrapers are categorized into four aspects based on their usage. These categories are explained below: [38]

- **Self-built or Pre-built:** Anyone can build their scrapers, but it requires advanced programming skills. Moreover, these scrapers can be updated to add some more features. In comparison, pre-build scrapers are the scrapers that are already created and available for download.
- **Browser Extension vs Software:** This categorization is based on how these scrapers are implemented. Browser extensions are extensions added to web browsers. Moreover, these are easy to run and embedded into the browser. Simultaneously, there is standard web scraping software installed on the system with some advanced features.

- **User Interface:** The web scrapers can have a user interface as a command-line interface or a graphical interface that provides a full-fledged interface.
- **Cloud vs Local:** A web scraper needs much computing power and resources to work on several URLs. A local web scraper runs on a system using the resources and network connection of the system. This will slow down the system due to high usage processing power and memory. So, a cloud-based web scraper runs on a cloud server and will take fewer resources.

Many web scraper tools and services are available online such as Mozenda, ParseHub, Diffbot and Import scraper. As discussed before in this chapter, web scraping has many advantages, like brand monitoring and price comparison. However, on the other side, some web scraping tools use malicious bots that cause harm to the websites. So, scraping should be performed only on the data allowed to scrape by the webmasters.

Chapter 3: Prevention of Data Harvesting

3.1 Detection of Data Harvesting

As mentioned in the chapter 2, “Analysis of Data harvesting”, data harvesting or web scraping using malicious bots may affect a company’s business in many ways. So, it is essential to prevent the data harvesting. To prevent the data harvesting, it is important to detect whether scraper bots are attacking the site or not. Some of the techniques used for detecting the presence of scraper bots on the websites are given below:

3.1.1 Content Duplication

When web-scraping is performed on a site, the site’s content is copied and published on another website without the site owner’s permission. This duplication of the original content shows that the presence of the web scraper bots on the site. The content duplication also proves that the site was affected due to the web scraping previously. Moreover, it will result in a poor SEO ranking of the website. [39]

3.1.2 Crawling Speed

The web scraping can be detected by examining the bot’s behaviour that shows an uneven number in metrics like page duration and page visits. Scraper bots are programmed to do repetitive tasks at very high speed in a short period to extract website content. So, based on the number of visits in seconds, scraper bots can be detected. [39]

3.1.3 Unwanted Spamming

Scraper bots can spam the website with unsolicited messages through the forms on the website. Many companies use these forms on their websites for registration and comments, making them face this problem. Some bots scour the web pages for information, then fill those pages with unsolicited messages. Due to this, genuine users will get interactions while using the website forms. [39]

3.1.4 Unusual Activity

Check for unusual activities such as continuous requests from the same IP address, which can be a sign of scraper bots. The scraper bots can also be detected by monitoring the number of searches in a short period.

The scraper bots can also be detected by analysing the user’s behaviour, such as mouse movements and how fast the user enters data into the form. Suppose a bot can enter the form in a fraction of seconds, but a real user will take a few minutes to fill the form. The bot can also

be detected by fingerprinting the browser's information such as the screen's resolution, time zone and browser type. [40]

3.1.5 Honeypot Webpage

The web scrapping can be detected by using the honeypot pages. In this method, a link invisible to the regular users is placed on the website's homepage. When a scraper bot tries to scrape all the pages on the website, it will click on all the links, including the invisible link. This will ensure that there is a presence of scraper bots. Then these bots are blocked from accessing the content of the pages. [40]

3.1.6 IP Tracking Tests

When a request is received, the server can notice several factors such as IP address, geo-location, Internet Service Provider (ISP) information, and connection type. This data should be analysed to know that the request is coming from a malicious bot or a genuine user.

Even though using these techniques, it is not easy to detect the scraper bot's presence without the help of an anti-bot solution because these bots are evolved to mimic human behaviour.

3.2 Best Practices for Prevention of Data Harvesting

Every organization needs to protect its data placed on their websites from data harvesting. Several prevention methods are developed to prevent the web scrapers bots and data harvesting. Some of the best practices are explained below.

3.2.1 CAPTCHA

"Completely Automated Public Turing Test to Tell Computers and Human Apart" (CAPTCHA) is the most effective method to prevent data harvesting. A challenge-response test is used to protect ad hoc search against bots by displaying a code and entering that code to determine whether the user is an actual human or a bot. [23]

This technique was first created in 1997 and thus requires the user to correctly evaluate and enter a sequence of the letters or numbers perceptible in a distorted image showed on the screen. If the entered data is correct, then the captcha system determines that the user is a human and allows them.

As we know that web scraping is performed with a bot's help and it is essential to prevent bots from accessing the website's content. This CAPTCHA system is beneficial for preventing bots.

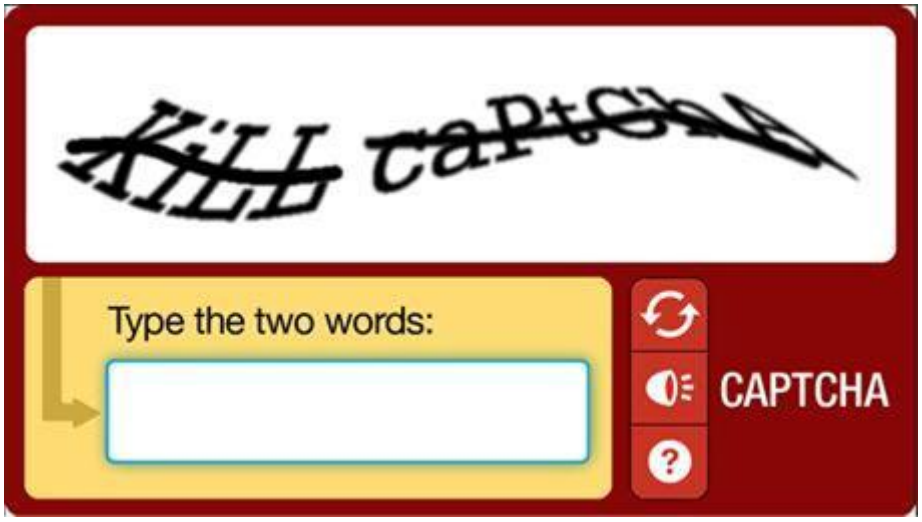

Figure 13 CAPTCHA [41]

Figure 13 shows the example of CAPTCHA. It displays the words “KiLL caPtChA” and asking the user to enter the exact two words in the text field. One important thing to remember while using CAPTCHA is every letter should be case sensitive. A human user can detect these two words and enter those into the text field. However, a bot cannot identify the exact words, so it fails in the CAPTCHA test. This testing will allow only real humans to proceed and prevent malicious bots from locating, crawling or auto completing the web forms.

Moreover, the CAPTCHA system can helps prevent spam due to malicious bots. Furthermore, the CAPTCHA is of many types such as “select the boxes with the traffic signals”, “tick the box”, and audio CAPTCHA for the visually impaired users” which can also help if the words are not visible properly.

However, more advanced scraper bots can use image and voice recognition systems to mimic human behaviour to bypass the CAPTCHA system. So, it is necessary to implement some more data harvesting prevention techniques simultaneously.

3.2.2 Change the Site’s HTML Markup Regularly

The HTML code of a website may have some vulnerabilities. If these vulnerabilities are not patched, there is a high chance of data harvesting due to scraper bots. So, the website’s HTML markup should be updated regularly.

The scraper bots also study the HTML markup patterns on the website and use these patterns to do scraping activities. It is good practice to regularly make changes in the HTML markup to confuse and discourage scraper bots from scraping the data. For example, HTML has an id attribute to identify an element. When the id attribute value is matched with the entered value,

then the action should be performed. By changing this attribute frequently, the bot's activities may be disrupted. [40]

3.2.3 Block IP Addresses

In most cases, the scraper bots use the same IP address while scraping the website's content. The IP addresses of users accessing the website should be monitored. If any scraper bot is detected, the bot's IP address should be blocked manually or based on the geolocation and Domain Name System Blacklist (DNSBL). [35]

The DNSBL (also known as Real-time Blacklist (RBL)) is a spam blocking list containing the user's IP addresses that cause spam. Moreover, it is based on the DNS system. When a server website gets requests from many IP addresses, the server then crosschecks each IP address of the user with the IP address in the blacklist. If the IP address is matched, then it should be blocked. [42]

3.2.4 Use Robots.txt File

Site owners use robots.txt files to permit or deny the scraper bots from traversing their websites. It is a standard used to guide the search engine bots to know which pages are allowed and not allowed to crawl. Every website has many webpages; some of them contain sensitive information. So, a robots.txt file can disallow the scraper bots from accessing these pages. However, this file informs the bots not to scrape, but it cannot prevent the bots from scraping the website. The bad bots can ignore the robots.txt file and scrape the web content.

For creating a robots.txt file, three steps are involved. They are explained below: [43]

1. The first step is to define the user-agent (software that acts on behalf of a user). There are several user agents available, so that any user agent can be defined at once. The following example shows the google bots as the user agent, which means that all the google services are allowed.

   User-agent: googlebot

   To grant access to all the user agents at once, a wildcard (*) is used, which is shown below:

   User-agent: *

2. After defining a user agent, then it is required to allow or disallow access to the subdirectories. If a subdirectory is required to provide access, then the directory's URL

path should be allowed. In the below example, the subdirectory "/search/about" can access by GoogleBot.

User-agent: googlebot

Allow: /search/about

3. If any particular subdirectory is blocked from access, then the URL path should be given below.

User-agent: googlebot

Disallow: /private/

It is possible to define multiple user agents at once. This can be done by simply writing one after another in the robots.txt file. An example for defining multiple user agents is given below.

User-agent: *

Disallow: /Search

User-agent: AdsBot-Google

Disallow: /maps/api/js/

Hence, using the robot.txt files will prevent scraping bots from scraping the contents from the pages mentioned as disallowed in the file.

3.2.5 Anti-Bot Solution

We know that scraper bots are required for data harvesting. And, it is essential to block malicious bots from scraping the website's content. So, an Anti-Bot solution or a Bot Manager plays a vital role in this prevention process. Many companies have been offering anti-bot and anti-scraping solutions for websites. These anti-bot solutions not only protect the website but also protect the APIs and mobile apps without affecting the performance.

An anti-bot solution is similar to the DLP solution, which prevents data leakage, whereas an anti-bot solution prevents the bots from scraping the data.

A typical anti-bot solution detects malicious bots by monitoring the website's activities for unusual behaviour. Moreover, they can analyse the visitor's interaction such as mouse movements, page traversals. However, this analysis is ineffective because most of the malicious bots mimic human behaviour. In this case, an advanced anti-bot solution can detect the non-human traffic that causes web scraping by using artificial intelligence and machine learning algorithms.

An anti-bot solution follows three steps for preventing the scraper bots from scraping the content. These steps are explained below:

- **Detection:** In this step, the anti-bot solution continuously monitors the traffic coming to the assigned websites or APIs without interfering with the users accessing the website. It uses many bot detection mechanisms such as monitoring unusual activity, crawling speed, fingerprinting (the process of accessing a device's software and connection attributes for creating a risk profile of it) and honeypot pages. Moreover, it can also differentiate the malicious bots and standard bots used by the search engines.
- **Prevention:** After detecting the malicious bots, the anti-bot service should mitigate these bots by using bot mitigation mechanisms such as blocking IP addresses, Implementing CAPTCHA and redirecting the bot traffic. They may also filter the bot traffic accessing the website.
- **Reporting and Analysis:** The anti-bot manager will log all the malicious bot's activities for future analysis. Moreover, report them to the administrator. These log files are then analysed to learn about the latest strategies used by the bots.

There are many anti-bot solutions available in the market, such as Radware Bot Manager and Akamai Bot Manager.

3.2.6 Terms of Use and Conditions

The implementation of terms of usage and conditions on the website's data will not prevent data harvesting. However, it may deter the attackers from harvesting the data. These terms of usage declare that the website's data should not be scraped without the site owner's permission. These terms also state that legal actions should be taken on the individuals when they harvest the data. [40]

Some sites use these terms to prevent the scrapers from scraping the website's content and placing it on another site. However, some sites may allow the scarper bots to scrape the data, but it is limited to personal use but not for commercial use.

Before scraping a webpage, the individual should read all the terms of use and conditions to avoid any legal actions.

Any one of the above measures will not protect the website from data harvesting, but all the prevention measures combining will safeguard the website's content.

Chapter 4: Conclusion

Data harvesting has been a concern for the website's owners since it started in the early 2000s. As the volume of data on the websites has been increasing tremendously, website security plays a vital role in protecting the websites from scraper tools and malicious bots. The scraper bots are improving their scraping techniques to bypass these prevention mechanisms to scrape the data. Moreover, there is a high possibility of developing new scraping tools and techniques in the future.

Although there are many ways to combat web scrapers, there is no mechanism that gives complete protection. So, it is essential to use more advanced technologies to protect the websites from data harvesting combined with the traditional prevention measures to make a reliable prevention system. This protection can be achieved by incorporating artificial intelligence and machine learning technologies into anti-bot solutions.

However, the same data harvesting prevention mechanisms or strategies will not be suitable for all companies. Every company has its own structure and data. It is essential to examine the organization's needs thoroughly before implementing a data harvesting security mechanism in that organization.

The data harvesting or web scraping techniques may evolve in the future, so webmasters should be prepared well to stop and mitigate data harvesting by developing and implementing new data harvesting prevention approaches to keep their user's data safely and securely.

References

[1] Abi Tyas Tunggal, "What is a Data Leak," UpGuard, Inc, 02 October 2020. [Online]. Available: https://www.upguard.com/blog/data-leak.

[2] J. D. Groot, "The History of Data Breaches," Digital Guardian, 24 October 2019. [Online]. Available: https://digitalguardian.com/blog/history-data-breaches.

[3] Dan Swinhoe, "The 15 biggest data breaches of the 21st century," CSO, 17 April 2020. [Online]. Available: https://www.csoonline.com/article/2130877/the-biggest-data-breaches-of-the-21st-century.html.

[4] Alix Postan, "The Yahoo Data Breach: What happened?," Uzado Inc, 11 October 2017. [Online]. Available: https://www.uzado.com/blog/the-yahoo-data-breach-what-happened.

[5] SelfKey Blog, "Facebook's Data Breaches – A Timeline," SelfKey Foundation, 11 March 2020. [Online]. Available: https://selfkey.org/facebooks-data-breaches-a-timeline/.

[6] CISOMAG, "Elasticsearch Server Exposed 1.2 Billion People Data," CISOMAG, 25 November 2019. [Online]. Available: https://cisomag.eccouncil.org/elasticsearch-server-exposed-1-2-billion-people-data/.

[7] Steve Symanovich, "Microsoft accidentally exposed 250 million customer records — What you should know," NortonLifeLock, [Online]. Available: https://www.lifelock.com/learn-data-breaches-microsoft-exposed-250-million-customer-records.html.

[8] IBM Security, "Cost of a Data Breach Report," IBM, 2020.

[9] P. Gordon, "Data Leakage - Threats and Mitigation," *SANS Institute Information Security Reading Room,* 2007.

[10] I. King, "Top 10 Causes of Data Leaks and How To Stop It From Happening," Nicebrains , 9 October 2017. [Online]. Available: https://nicebrains.com/technology/data-leak-causes/.

[11] D. Gibson, CompTIA Security+ Get Certified Get Ahead SY0-501 Study Guide, 2017.

[12] Lithmee, "What is the Difference Between Data Loss Prevention and Data Leakage Prevention," PEDIAA, 5 August 2019. [Online]. Available: https://pediaa.com/what-is-the-difference-between-data-loss-prevention-and-data-leakage-prevention/#Data%20Leakage%20Prevention.

[13] Y. E. L. R. Asaf Shabtai, A Survey of Data Leakage Detection and Prevention Solutions, Springer, 2012.

[14] E. Bickerstaffe, "Data Leakage Prevention," Information Security Forum Limited, 2018.

[15] ORION CASSETTO, "Data Loss Prevention Policy Template," Exabeam, 2 May 2019. [Online]. Available: https://www.exabeam.com/dlp/data-loss-prevention-policy-template/.

[16] McAfee, "What Is DLP and How Does It Work?," McAfee, [Online]. Available: https://www.mcafee.com/enterprise/en-us/security-awareness/data-protection/how-data-loss-prevention-dlp-technology-works.html#overview.

[17] McAfee, "McAfee Data Loss Prevention 10.0.500 Product Guide (McAfee ePolicy Orchestrator)," McAfee, 25 July 2017. [Online]. Available: https://docs.mcafee.com/bundle/data-loss-prevention-10.0.500-product-guide-epolicy-orchestrator/page/GUID-EC4B158C-C7AC-4F1F-8335-D63A692B7544.html.

[18] Jeff Melnick, "10 Best Practices Essential for Your Data Loss Prevention (DLP) Policy," Netwrix, 16 July 2019. [Online]. Available: https://blog.netwrix.com/2019/07/16/10-best-practices-essential-for-your-data-loss-prevention-dlp-policy/.

[19] J. Dion, *CompTIA Security+ (Study Notes),* DionTraining, 2020.

[20] L. Rogers, *INTERNET SECURITY-MINT 712,* University of Alberta, 2015.

[21] M. Bonner, "Why Your Business Needs a Data Breach Response Plan," The balance small business, 28 January 2020. [Online]. Available: https://www.thebalancesmb.com/data-breach-response-plan-for-your-business-4154834.

[22] Yaniv Masjedi, "The CIO's Data Breach Response Plan for 2020," Nextiva, 15 February 2019. [Online]. Available: https://www.nextiva.com/blog/data-breach-response-plan.html.

[23] Caspio, Inc., "What You Need to Know about Data Harvesting and How to Prevent it," Caspio, Inc., 27 August 2014. [Online]. Available: https://blog.caspio.com/what-you-need-to-know-about-data-harvesting-and-how-to-prevent-it/.

[24] Marco Tapia, "Data Harvesting: What you need to know.," PicNet Pty Ltd, [Online]. Available: https://picnet.com.au/blog/data-harvesting-need-know/.

[25] Salient Marketing, "The first web robot – Wanderer," Salient Marketing, [Online]. Available: https://salientmarketing.com/resources/the-first-web-robot-1993/.

[26] DataDome, "Web scraping protection: How to protect your website against crawler and scraper bots," DataDome, [Online]. Available: https://datadome.co/bot-management-protection/scraper-crawler-bots-how-to-protect-your-website-against-intensive-scraping/.

[27] D. Parrack, "Facebook Addresses the Cambridge Analytica Scandal," makeuseof, 22 March 2018. [Online]. Available: https://www.makeuseof.com/tag/facebook-cambridge-analytica-scandal/.

[28] Zeljka Zorz, "Sensitive data on 198 million US voters exposed online," Help Net Security, 19 June 2017. [Online]. Available: https://www.helpnetsecurity.com/2017/06/19/us-voters-data-leak/.

[29] M. Ellis, "10 Real Examples of When Data Harvesting Exposed Your Personal Info," makeuseof, 24 May 2018. [Online]. Available: https://www.makeuseof.com/tag/data-harvesting-personal-info/.

[30] Harkiran, "What is Web Scraping and How to Use It?," GeeksforGeeks, 22 June 2020. [Online]. Available: https://www.geeksforgeeks.org/what-is-web-scraping-and-how-to-use-it/.

[31] Radware, "The Difference Between Good Bots and Bad Bots," Radware, [Online]. Available: https://www.radwarebotmanager.com/good-bots-and-bad-bots/.

[32] F. A, "What is Data Harvesting and How To Prevent It," WebScraping, 18 February 2020. [Online]. Available: https://wscraper.com/what-is-data-harvesting-and-how-to-prevent-it/.

[33] A. Rao, "Introduction to Web Scraping," GeeksforGeeks, 6 November 2019. [Online]. Available: https://www.geeksforgeeks.org/introduction-to-web-scraping/?ref=rp.

[34] Radware Bot Manager , "An Overview of Web Scraping Techniques," Radware, [Online]. Available: https://www.radwarebotmanager.com/What-are-the-Different-Scraping-Techniques/.

[35] "Web scraping," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Web_scraping#Techniques.

[36] Radware , "The Anti-scraping solution for your website & app," Radware, [Online]. Available: https://www.radwarebotmanager.com/anti-scraping/.

[37] Discover.bot, "Web Scraping with Bots: The Good and Bad," Discover.bot, 2 October 2018. [Online]. Available: https://discover.bot/bot-talk/web-scraping-with-bots-the-good-and-bad/.

[38] M. Perez, "What is Web Scraping and What is it Used For?," Parsehub, 6 August 2019. [Online]. Available: https://www.parsehub.com/blog/what-is-web-scraping/.

[39] Radware, "3 common indications your website is overrun by scraper bots," Radware, [Online]. Available: https://www.radwarebotmanager.com/indications-your-website-is-overrun-by-scraper-bots/.

[40] Sachin, "Web Scraping Prevention 101 – How To Prevent Website Scrapping," Techicy, 25 May 2020. [Online]. Available: https://www.techicy.com/web-scraping-prevention-101-how-to-prevent-website-scrapping.html.

[41] Global Accessibility News, "Advocacy group wants Google to rethink CAPTCHA," Global Accessibility News, [Online]. Available: https://globalaccessibilitynews.com/2014/02/06/advocacy-group-wants-google-to-rethink-captcha/.

[42] DNSBL.info, "What is a DNSBL?," DNSBL.info, [Online]. Available: https://www.dnsbl.info/.

[43] Wikipedia, "Robots exclusion standard," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Robots_exclusion_standard.

[44] R. Tahboub and Y. Saleh, "Data Leakage/Loss Prevention Systems (DLP)," *International Journal of Information Systems,* vol. 1, 2014.

[45] L. C. F. L. D. (. Yao, "Enterprise data breach: causes, challenges, prevention, and future directions," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 2017.