7-2020

EMAIL DATA BREACH ANALYSIS AND PREVENTION USING HOOK AND EYE SYSTEM

Shubhankar Jayant Jathar
California State University - San Bernardino

EMAIL DATA BREACH ANALYSIS AND PREVENTION USING HOOK AND EYE SYSTEM

A Project

Presented to the

Faculty of

California State University,

San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Information Systems and Technology

by

SHUBHANKAR JAYANT JATHAR

2020

Email Data breach analysis and prevention using Hook and Eye System

A Project

Presented to the

Faculty of

California State University,

San Bernardino

by

Shubhankar Jayant Jathar

July 2020

Approved by:

Vincent Nestler, Committee Chair, Information & Decision Sciences

Conrad Shayo, Committee Member

Jay Varzandeh, Committee Member

ABSTRACT

Since the recent COVID-19 outbreak, there have been reports of high incidents of cybercrime in the health sector. This project attempted to analyze the different types of breaches that are found online and are practiced to steal valuable information, specifically focusing on the health sector. The analysis found that most data breaches are done through emails. Thus, this project designed a system that will help in reducing the risk of getting hacked through email. A web-based portal is created using a machine learning model which identifies a suspicious email before it gets delivered to the receiver.

The project used the XAMPP server and HTML, CSS language to develop the website. Microsoft Excel and Tableau are used for statistical analysis and graphical representation.The results from the use of the prototype developed in this project are as follows: From the analysis of the data breach incidents it was inferred that most of the data breaches cause due to phishing attacks. (From analysis) 34% of the total data breach occur due to phishing attacks. The category of the phishing attacks include Remote Access, W-2 Scam, Ransomware and Automated Information Exfilteration. The system developed would eventually lead to deep email filtering which will prevent phishing and other means of data breaches form email. It will increase the efficiency of the system as the filtering will be detailed and will involve machine learning. HKNI system is expected to filter 100% of the emails received and would deliver clean emails and terminate the bad ones.

CONTENTS

CHAPTER ONE

INTRODUCTION

A data breach is the intentional or unplanned vulnerability of confidential information to an unauthorized person or organization. The data is the most important asset as it has a lot more worth than anything. It also causes a threat to the person, organization, or entity if the data is leaked. A data leak can cause a great loss of money, integrity, image, and reputation of an individual, group, or organization. The role of a cybersecurity specialist is to mitigate the probability of a data breach. Data breach rates in most organizations are two to three attacks per week (Forbes, 2019). A lot of efforts are taken to reduce data breach incidents but every time the attacker is successful due to ease of technology available in the market and on the dark web. It is also very difficult for cyber police to shut down the dark web from which many attackers try to invent and share new techniques.

After the analysis, it has been seen that most of the data breach which takes place is due to phishing attempts which is mainly due to emails. Nowadays, emails are the primary way of communication when it comes to writing and there is a need to increase security. Some of the incidents like the one which happened with the email.it organization where 600,000 emails were leaked and put up on sale on the dark web (Amer Owaida, April 2020). The email which is received is the same and will not be able to tell the difference in the phishing email and normal email. I will be showing an example of such emails further.

PROBLEM STATEMENT

The current data breach report suggests that most of the breach incidents reports are due to email and in this project, the discussion is about the solution to decreasing the attempts of a data breach due to emails.

This project focuses on the following things.

1. Analyzing the main cause of data breach incidents.
2. Identifying the main causes behind e-mail enabled data breaches.
3. Design and implement a web-enabled system solution that will mitigate the problem by making sure that organizational email is filtered before reaching its destination.

CHAPTER TWO

LITERATURE REVIEW

The HKNI system mentioned in the project is a modified version of the present system which increases the security in the email breaches. There was no direct research related to the HKNI system. Some present systems are used by Google called G Suite which does the same work but the HKNI has a built-in defense system that is not mentioned in the G Suite program. Studies show that the HKNI system can cause a great change in the field of cybersecurity as there is a consistent increase in data breaches every day.

A recent study has shown that data breach incidents cause a big threat to the stock market which eventually causes economic crisis (Campbell et al. 2003; Cavusoglu et al. 2004; Acquisti et al. 2006; Kannan et al. 2007; Gordon et al. 2011), it is proposed that there should be spread of knowledge about what steps should be taken to prevent data breaches (Belsis et al. 2005). There have been efforts taken to reduce cybercrime by spreading consumer education. There is a special team called US Computer Emergency Readiness Team (US-CERT) (CERT, 2008) which offers an immediate help response on all types of cyberattacks and incidents using various exploits, cybersecurity alerts, and vulnerabilities. Almost 45 states have passed a bill where they are required to notify any breach taking place. Additionally, the Data Accountability and Trust Act being considered by Congress would give a uniform national notice strategy (Pike 2009).

Taking the reference to a machine learning-based cyber threat detecting model so that the top threats can be identified, using a modified version of the machine learning model in the system (M.T. Khorshed, 2012). A mechanized system to react to digital dangers with proper reaction plans is proposed. It works by coordinating and assessing operational, money related, and threat impact models (G. Gonzalez-Granadillo, 2018). A system was proposed that would learn which words and links are harmful in the email and automatically add them to the database to get checked in the system. In one of the projects, it was shown that hacking is a decreasing trend but this project proves that hacking has an increasing graph and it is developing as the technology advances (Shu, X., Tian, K., Ciambrone, A., Yao, D. 2017). Smith (2016) focused on data breaks on social insurance associations to decide the relationship between information security breaches, data storage areas, business partners, secured elements, and several people influenced. It was found that 70% of breaches include human services suppliers and security episodes frequently comprised of electronic or other computerized data.

CHAPTER THREE

THE METHODOLOGY

The first part includes the definition of all the tools and keywords included in the project. Different strategies are mentioned which are used to achieve success. Data has been analyzed according to the data which was downloaded from the Kaggle and different websites. This clearly states the levels of information affectability, techniques utilized in preventing information breaks.

The second part will mainly focus on designing a system to prevent email phishing – Hook and Eye System, designing a web-based portal to test the system.

The third part will address the solution for organizations, how these breaks can be diminished, and what arrangements can be applied by each state depending on the urgent components. This will be accomplished by gathering information from different sources like the CIA and Kaggle.com. Starting now and into the near future, analysis of the data and propose a likely responsible for the troublesome request.

Data Sources and from where they are collected

The information utilized in this examination was acquired from different sources and secured in the period somewhere in the range of 2000 and 2019. Attributes of information breaches were acquired from breaking down the accessible information. The information was gathered, cleaned, controlled,

arranged, and afterward broken down. Most highlights were gotten for the period somewhere in the range of 2000 and 2019, including the element assaulted, people influenced, the expense for recovery, state's influences, year in which the occurrence occurred, kind of breach, sort of associations.

Data Set Description

An informational index (or dataset) is an assortment of information. On account of even information, an informational collection relates to at least one database table, where each segment of a table speaks to a specific variable, and each line compares to a given record of the informational index being referred to. The informational collection records relate to every one of the factors, for example, the tallness and weight of an item, for every individual from the informational collection. Each worth is known as a datum. Informational collections can likewise comprise an assortment of records or documents.

Dataset Definitions

entity	Name of the organization which were affected by data breaches
state	The states in the US which were affected by a data breach
Individual_affected	An individual who was targeted for hacking or whose data was leaked

Breach_submission	The date on which the incident was reported
Location_of_brech	The location from which the breach was initiated
Type_of_breach	The type of hacking method used to infiltrate the data
Record_lost	Number of records lost during the breach
Year	The year in which the data breach took place
Sector	The sector to which organization belongs to
Method	The process behind the data breach
Data_sensitivity	The level of the importance of the file which was leaked in the data breach

Tools

Web Crawler

A web crawler is a program or mechanized content that peruses the World Wide Web in an orderly, computerized way. This procedure is called Web creeping or spidering. Many real locales, specifically web crawlers, use spidering as a method for giving modern information. Web crawlers are for the most part used to duplicate all the visited pages for later handling by a web crawler, that will record the downloaded pages to give quick quests. Crawlers can likewise be utilized for

robotizing support assignments on a Web webpage, for example, checking joins or approving HTML code.

Web Scrapper

Web scraping is the process of using bots to extract content and data from a website. Unlike screen scraping, which only copies pixels displayed onscreen, web scraping extracts underlying HTML code, and, with it, data stored in a database. The scraper can then replicate the entire website content elsewhere. Web scraping is used in a variety of digital businesses that rely on data harvesting.

XAMPP Server

XAMPP is a contraction for cross-stage which consists of Apache, MySQL, PHP, and Perl, and it permits to construct an offline website, on a web server on users' PC. This basic and lightweight arrangement deals with Windows, Linux, and Mac – thus the "cross-stage" part. Since WordPress isn't an independent application, XAMPP gives two basic segments to its establishment – Apache, which is utilized to make the neighborhood server and MySQL which can be used as a database for the site.

Tableau

Scene programming is one of the quickest developing information representation devices which is now being utilized in the BI business. It is the most ideal approach to change for the crude arrangement of information into an effectively reasonable configuration with zero specialized abilities and coding

information. Tableau is a ground-breaking and quickest developing information representation device utilized in the Business Intelligence Industry. It helps in disentangling crude information into the effectively justifiable configuration. Information examination is quick with Tableau and the representations made are as dashboards and worksheets.

Why Tableau?

Tableau is enormously utilized because information can be broken down rapidly with it. Likewise, representations are created as dashboards and worksheets. Scene permits one to make dashboards that give noteworthy bits of knowledge and drives the business forward. Scene items consistently work in virtualized situations when they are arranged with the correct basic working framework and equipment. Tableau is utilized to investigate information with a boundless visual examination.

Traditional Method	Tableau
Prior programming skills	No programming skills required
Focused on only one type of database	Combines different types of database spreadsheets, databases, cloud data, and even big data such as Hadoop
Decision-makers must ask the IT people to retrieve any information from the database	Decision-makers can directly use the dashboard to retrieve any information from the database
Mostly depends on Query languages	The query is done behind the scene
Combining different types of the database is difficult	Different types of databases can be combined easily
Not every business intelligence tool offers an interactive dashboard	The interactive dashboard is easy to build, and it makes data visualization quick and efficient
Mostly designed for large businesses	Perfect BI solution for small, medium, and large businesses, and even for non-profits
Comparatively expensive	Comparatively affordable
Time-consuming	Timesaving

Table 1:Comparison of Traditional Method and Tableau (Rahman, 2015)

Data Cleaning

Information purifying or information cleaning is the adjusting (or evacuating) degenerate or wrong records from a recordset, table, or database and alludes. Distinguishing fragmented, changing, or erasing the filthy or coarse data. Data

purging might be performed intuitively with information fighting devices, or as bunch handling through scripting.

There are many anomalies in the dataset which include

1. Mismatched Columns
2. If the day is missing
3. Duplicate Records
4. Null records
5. Different date formats.

All the anomalies should be cleared to represent the data without any errors. So the process for data cleansing is as follows.

- Mismatched Columns

  Some of the time to change the granularity of certain information, either to decrease the measure of information created from the stream or to adjust information to other information should consolidate or association. For instance, this should total deal information by the client before getting a business table together with a client table.

  To change the granularity, utilize the Add Aggregate alternative to make a stage to total or gathering information. Regardless of whether the information is accumulated or assembled relies upon the information type (string, number, or date).

- If the day is missing

Turn on "Show Missing Values" for dates and canisters. (Right snap or utilize the drop-down menu on the field in the view and select Show Missing Values). Since Tableau knows the min/max esteems and augmentations for a date or container that characterizes a Row/Column header, it can fill in missing qualities and therefore give what's absent.

- Duplicate Records

  It is important to realize which measurement in the information source is extraordinary for each line esteem or the mix of measurements. For instance, if the measures on Table A have a one of a kind line identifier dependent on Date/Time, utilize that measurement to expel copy esteems.

- Null Records

  At the point when the measure is dragged or persistent date to the view, the qualities are appeared along with a nonstop pivot. On the off chance that the field contains invalid qualities, or if there are zeroes or negative qualities on a logarithmic pivot, Tableau can't plot them. An invalid worth is a field that is clear and implies absent or obscure qualities. At the point when these qualities exist, Tableau shows a marker in the lower right corner of the view that says obscure qualities exist. Snap the marker and look over the accompanying alternatives:

  - Channel Data - prohibit the invalid qualities from the view utilizing a channel. At the point when information is searche, the invalid

qualities are additionally prohibited from any estimations utilized in the view.

- Show Data at Default Position - show the information at a default area on the pivot. The invalid qualities will, in any case, be remembered for figurings. The default position relies upon the information type. The table underneath characterizes the defaults.

- Mismatch column

  Right, when data is imported in Tableau, it normally perceives the data kind of the area; in any case, in our dataset scene, it couldn't discover a bit of the fragment. As a component of the data cleaning process, these segment datatypes were updated.

Data Sensitivity

Sensitive data cannot avoid being information that must be guaranteed against unapproved get to. Access to delicate data should be confined through satisfactory data security and information security practices proposed to prevent unapproved disclosure and data breaks. Affiliation may need to make sure about unstable data for good or real necessities, singular insurance, regulatory reasons, trade insider realities, and other essential business information. Such data could introduce extended social, reputational, legal, employability, or security risk for just

as customers at whatever point revealed and is consistently the goal of corporate spying. Pair this with the climb of managerial assessment for certain organizations and it has indeed a necessity for data the load up, vendor peril the load up, untouchable risk the administrators, and advanced security than whenever in ongoing memory. The hardship, misuse, change, or unapproved access to the most fragile data can hurt the business, ruin customer trust, break customer insurance and in uncommon cases, impact the security and overall relations of nations.

The underlying advance is to develop a data portrayal technique to describe sensitive data and set up rules for its confirmation. Even though this report is the purpose behind ensuring that fragile data is dealt with reasonably, various data request plans to come up short for various reasons, for instance, going with:

- The technique uses complex language that is difficult for delegates to appreciate and follow, leaving them with a more prominent number of requests than answers.
- It isn't supported through getting ready and doesn't fit into the affiliation's work forms. Its destinations are unnecessarily forceful and difficult to achieve.
- It fails to outline the obligations taking everything into account.
- It doesn't convince laborers concerning the criticalness of data gathering.
- Policies are created once and never explored and refined.

As affiliations advance and business needs change, a data portrayal procedure can get unimportant inside scarcely any years.

CHAPTER FOUR

DATA ANALYSIS AND VISUALIZATION

A glance of US database Incidents

Number of Incidents

As seen in recent years, the US Data Breach incidents have drastically increased due to advancements in technology. Figure 4.1 shows the incidents from the year 2003 to 2019.

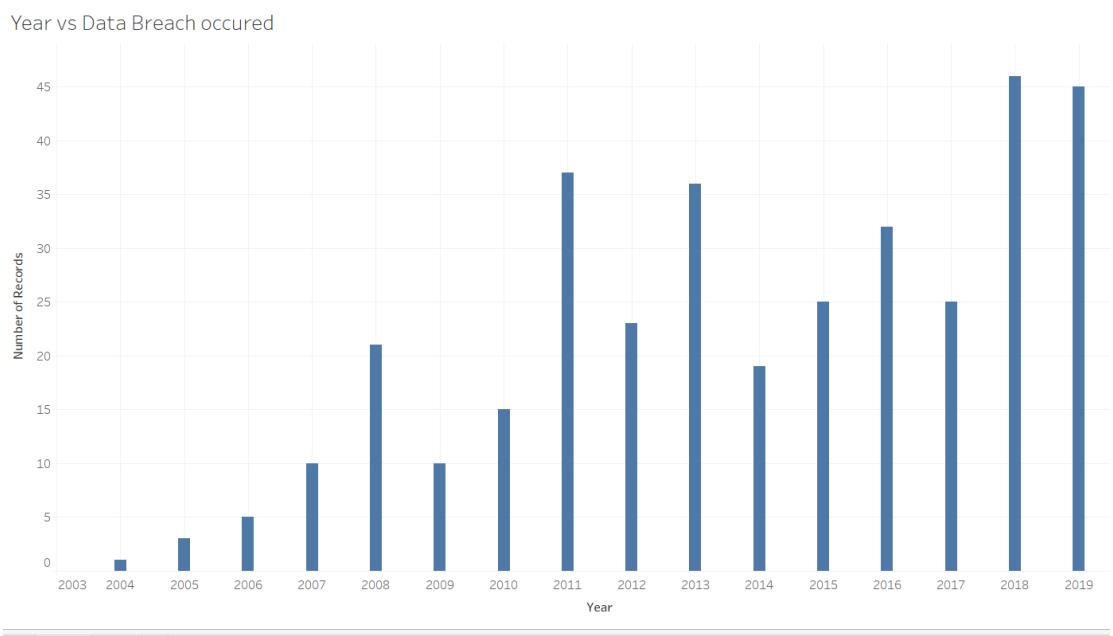

Figure 4 1:Incident Trend Over the Years

Number of incidents vs States in the US

The figure 4.2 shows the number of incidents by each state. A range of colors is used to show how the number goes from increasing from a state to another.

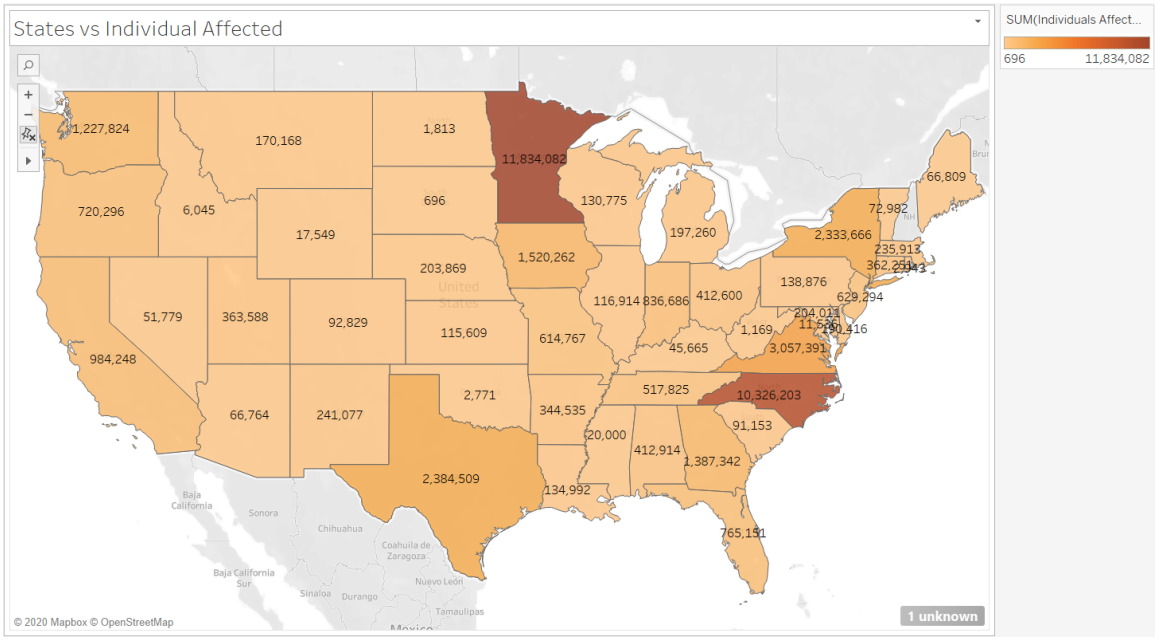

Figure 4 2:Individual affected in Each State

The top five states in which most of the individuals were affected are shown below in figure 4.3.

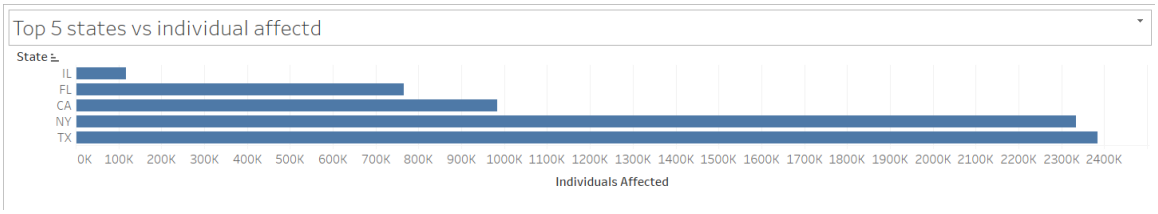

Figure 4 3:Top Five States with a High Number of Incidents

Incidents By States

The figure 4.4 shows visualization of several incidents happening by longitude and latitude. There are more incidents on both the east and west coast as compared to the central part. Texas shows a large number of incidents.

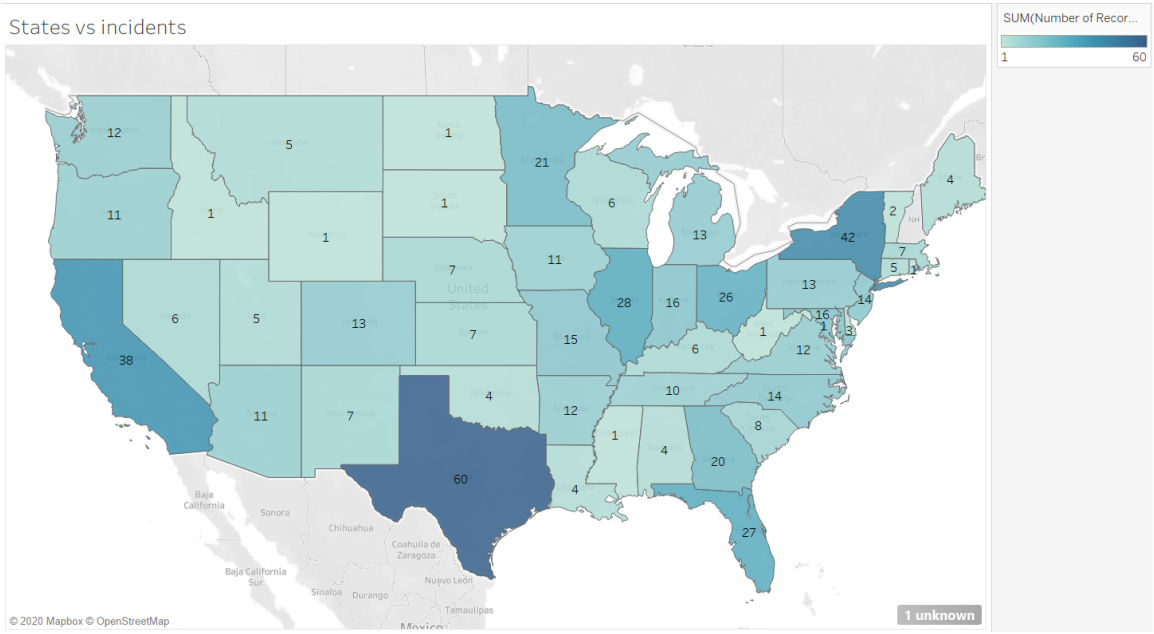

Figure 4 4:Incidents by Sates

The figure 4.5 shows the top 10 states where the data breach incidents took place. Figure 4.5 shows the descending order of the states with their count.

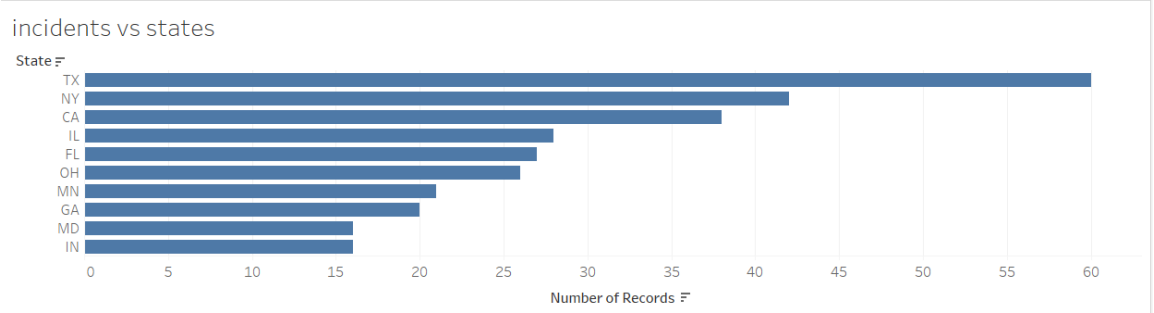

Figure 4 5:Top 10 states vs the incidents

Incidents occurred vs Months

Figure 4.6 shows individual affected by month from year 2018 to 2020. Figure 4.6 displays no trend as there is no increasing or decreasing trend. The month of July in year 2019 shows a sudden increase in the number of individuals affected.

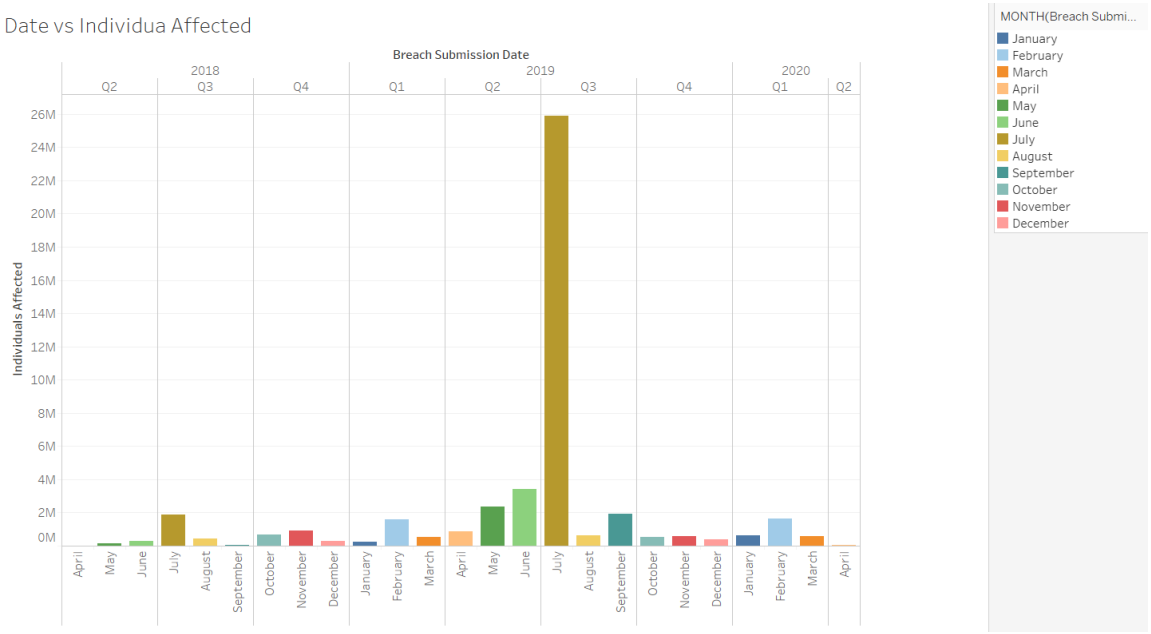

Figure 4 6:Individual affected by Month

Incidents occurred vs Day

Between 2018 and 2019, most of the incidents took place during weekends (Friday, Saturday, Sunday). Relatively, there are fewer incidents on weekdays (Monday to Thursday)

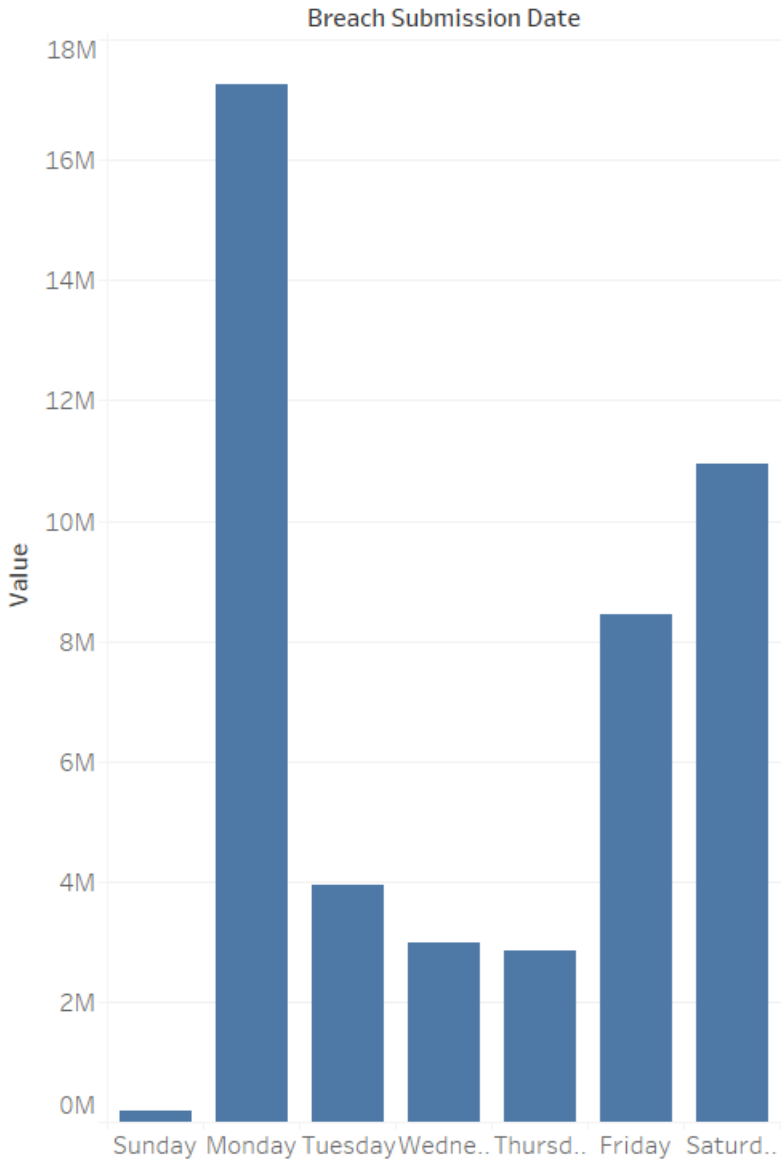

Figure 4 7: Individual affected by Day

Incidents vs Entity

Between 2000 and 2019, there was a large number of organizations were affected. There were several factors for the attack. I will be mentioned all the

factors which lead to the data breach. This graph in Figure 4.8 shows the entity affected vs the individual affected.

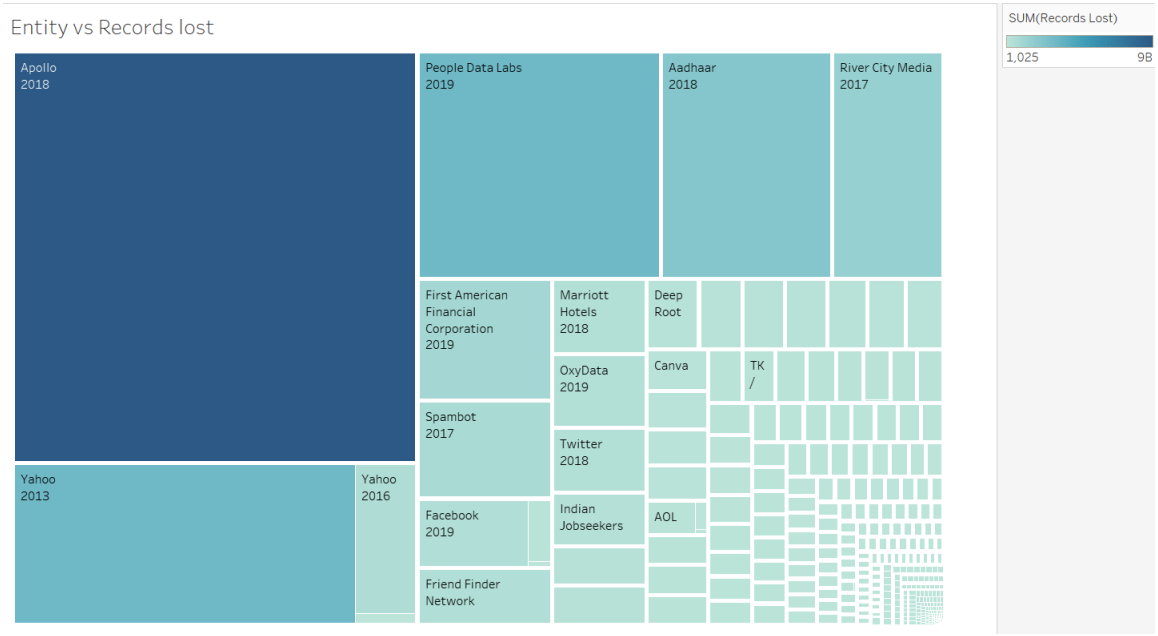

Figure 4 8:Incidents individual affected vs entity affected.

Sector vs Record lost

The Figure 4.9 below shows the sectors which are affected by the data breach and also how many records are lost per sector.

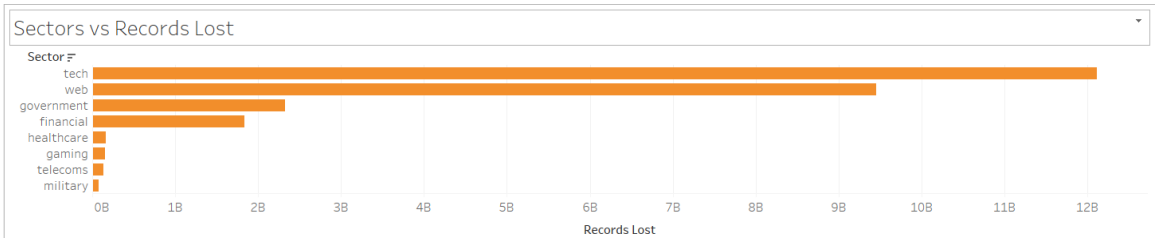

Figure 4 9: Sector vs Record Lost

In this graph, it can be can seen that the records lost represent the technical part. Most of the data which is lost is tech-related and the top 5 sectors affected are Tech, web-based, government, financial, healthcare.

Hacking Methods

There are several hacking methods used by hackers to infiltrate into the organization. The graph in Figure 4.10 below shows the 4 different types of methods of hacking and also the count of the records lost.

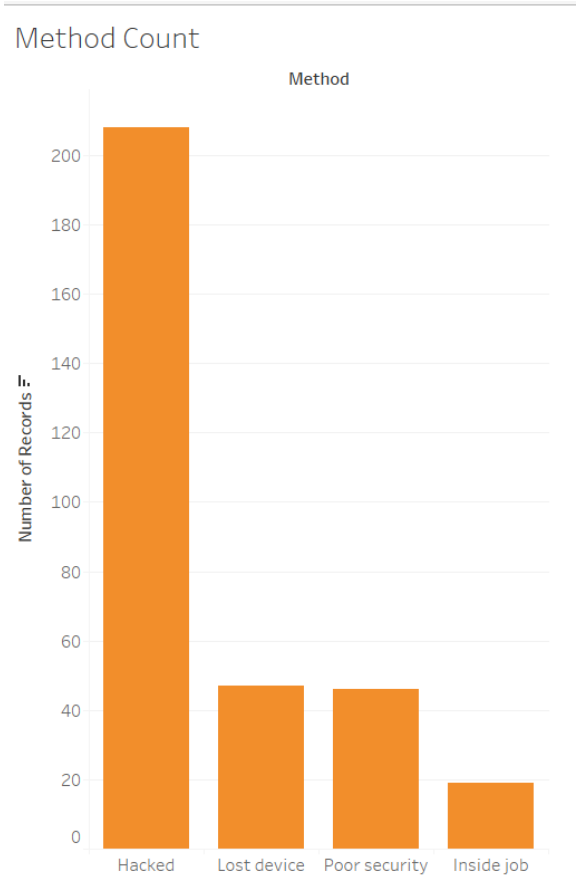

Figure 4 10:Method vs Number of records lost

Data Sensitivity

The graph in the Figure 4.11 below shows the list of files affected vs the data sensitivity. Every file has separate sensitivity according to the level of the file according to the importance of the file.

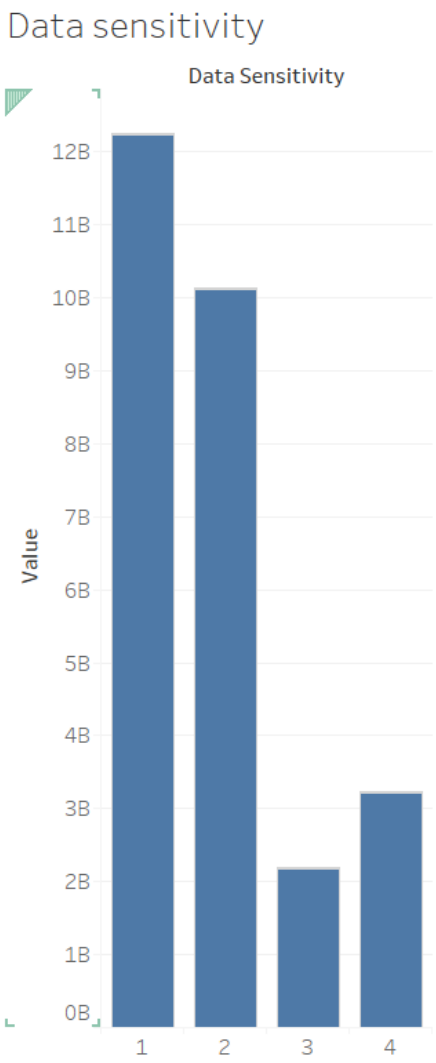

Figure 4 11:Nunber of records lost by Sensitivity

Year vs Record lost

In the below graph of Figure 4.12, all the different sectors are mentioned and the year in which they are affected is also mentioned below.

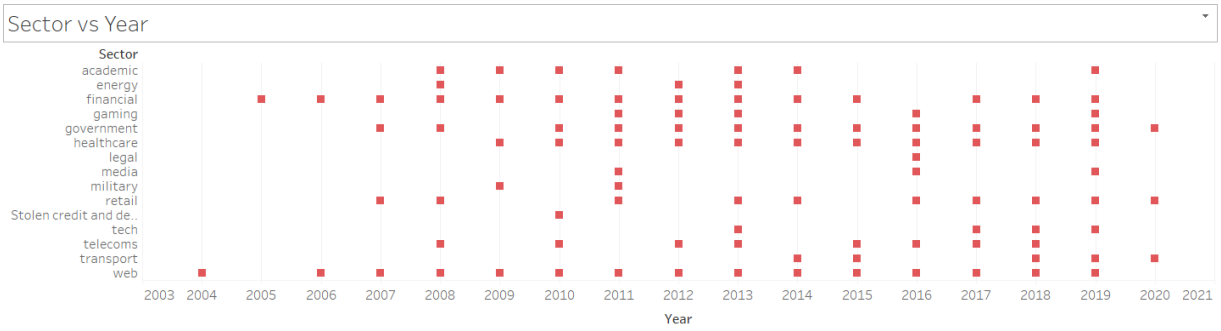

Figure 4 12: Sector Affected vs year

In Figure 4.12 above it can be seen that the attacks are increased from the year 2004 to 2020.

Incidents by Type

There are different methods for data breaches. These methods vary from sector to sector. The graph for the incidents by type of breach is shown below in Figure 4.13

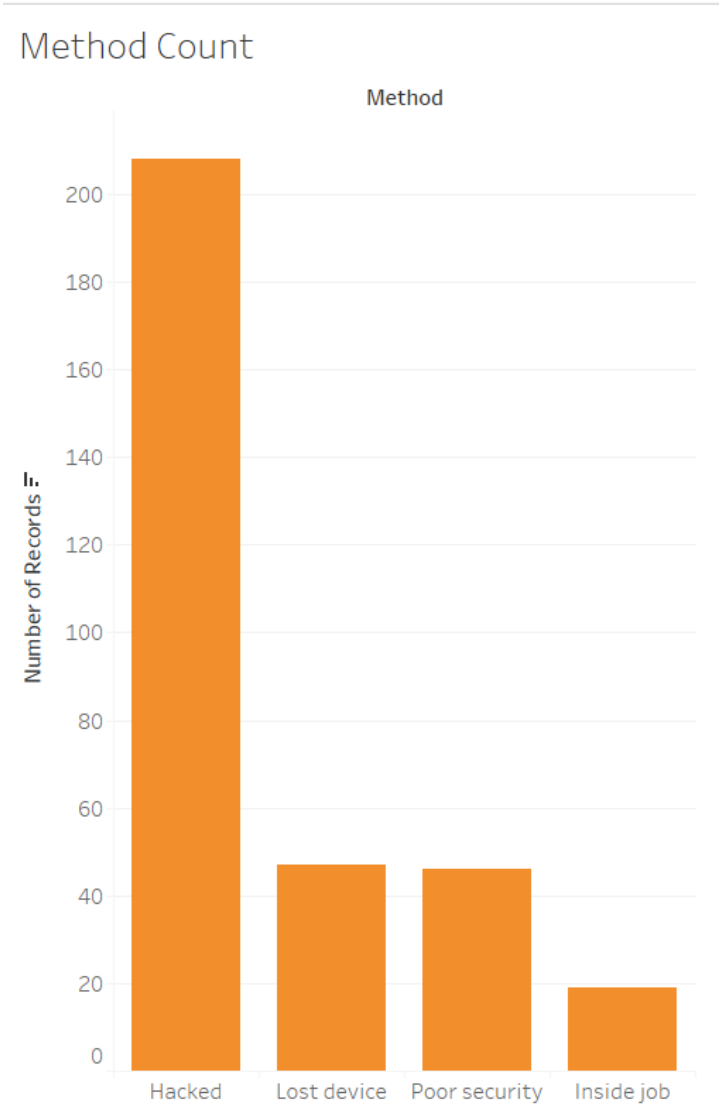

Figure 4 13:Incendents by methods

Data breach through location

The medium through which data breach occurs mostly matters as its the main cause of data breach. There are several ways or the medium which causes the breach. The following Figure 4.14 shows it.

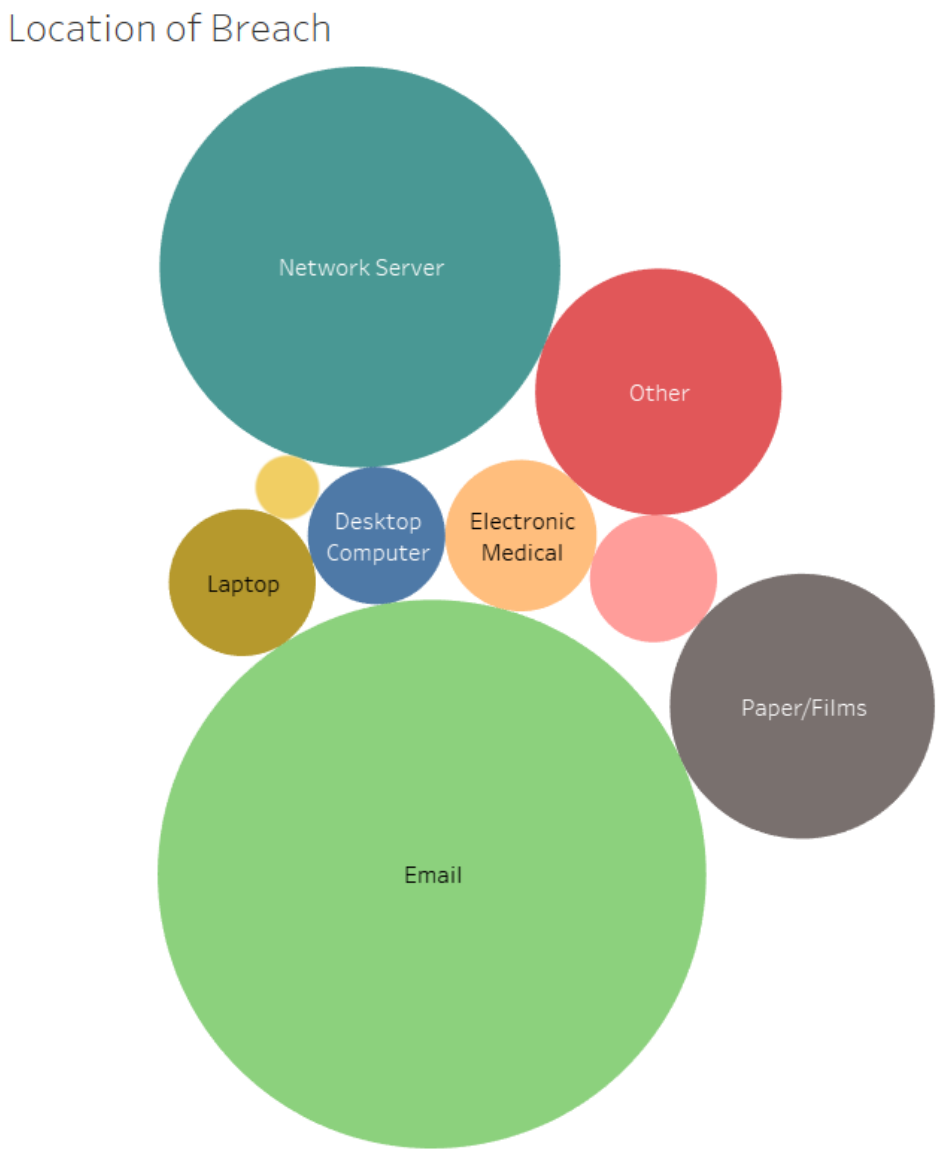

Figure 4 14:Data breach by type.

Tactics used in breaching

The below graph shows various tactics used in data breaching.

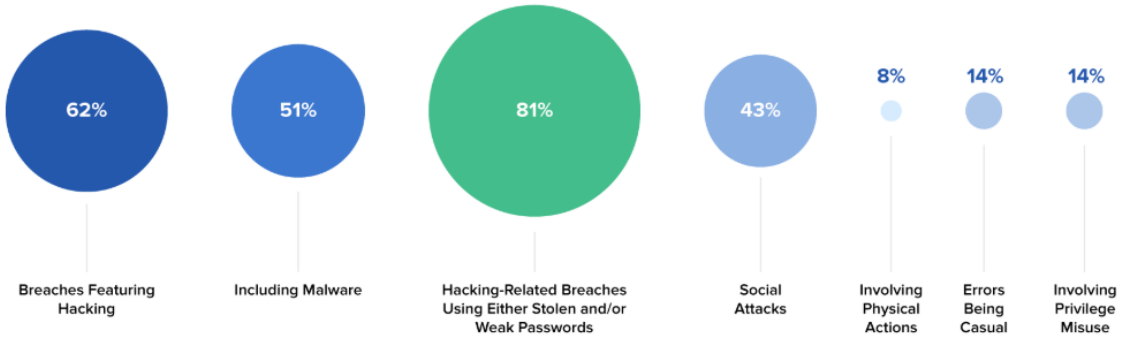

Figure 4 15: Tactics by percentage

In this Figure 4.15, it clearly states that the breach occurs because of hacking-related method and using either stolen and weak passwords. From this, I analyzed that the breach occurs because of phishing for which the main cause is emailed.

Data breach Incidents Description by Word Cloud

A 'word cloud' is a visual portrayal of word recurrence. The more ordinarily the term shows up inside the content being broke down, the bigger the word shows up in the picture produced. Word mists are progressively being utilized as a basic device to distinguish the focal point of composed material. They have been utilized in legislative issues, business, and instruction, for instance, to envision the substance of political discourses.

Based on the incident of the data breach and the factors leading to the data breach, this word cloud has been generated which includes the methods, factors, and main cause for data breaches.

Figure 4 16:Databreach Description Word Cloud

Number of the data breach and exposed records

The measurement presents the advancement of digital assaults after some time. It presents the recorded number of information breaks and records uncovered in the United States somewhere in the range of 2005 and 2014. In 2014, the number of information breaks in the United States added up to 783 with more than 85.61 million records uncovered.

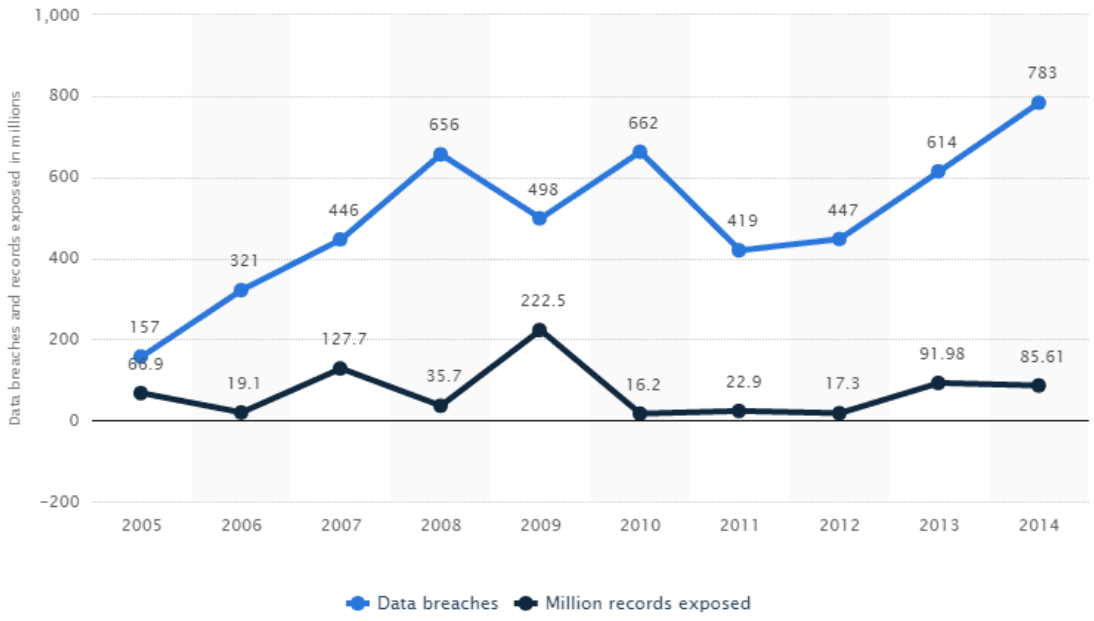

Figure 4 17:Number of the data breach and exposed records (2005-2014)

Data breach by sector

This Figure 4.18 shows that the main data breach is occurred because of email as Yahoo is one of the biggest provider for email communications.

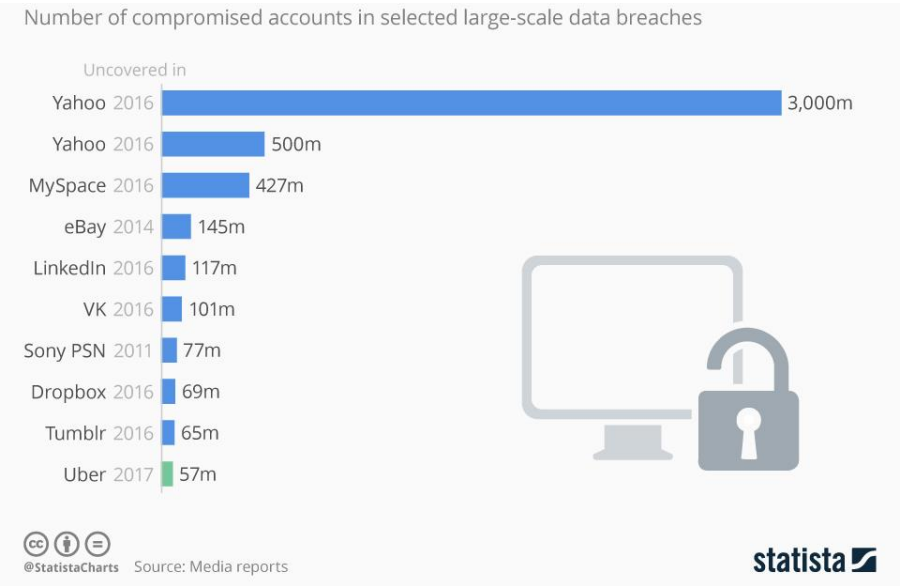

Figure 4 18:Sector vs data breach

Data breachers according to age.

The following Figure 4.19 shows the age of hackers by percentage.

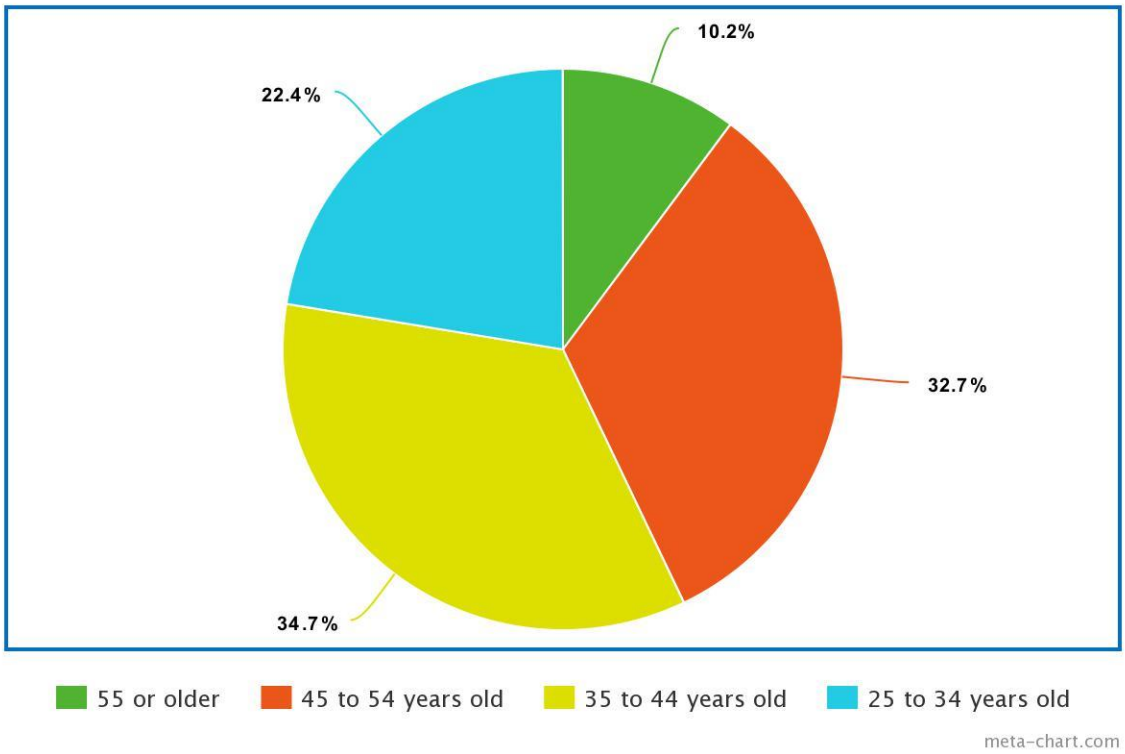

Figure 4 19:Hackers by Age Group (2004-2018)

The main cause for data breach

So after all the analysis and graphs, it was concluded that the main factor which is responsible for a data breach is through emails.

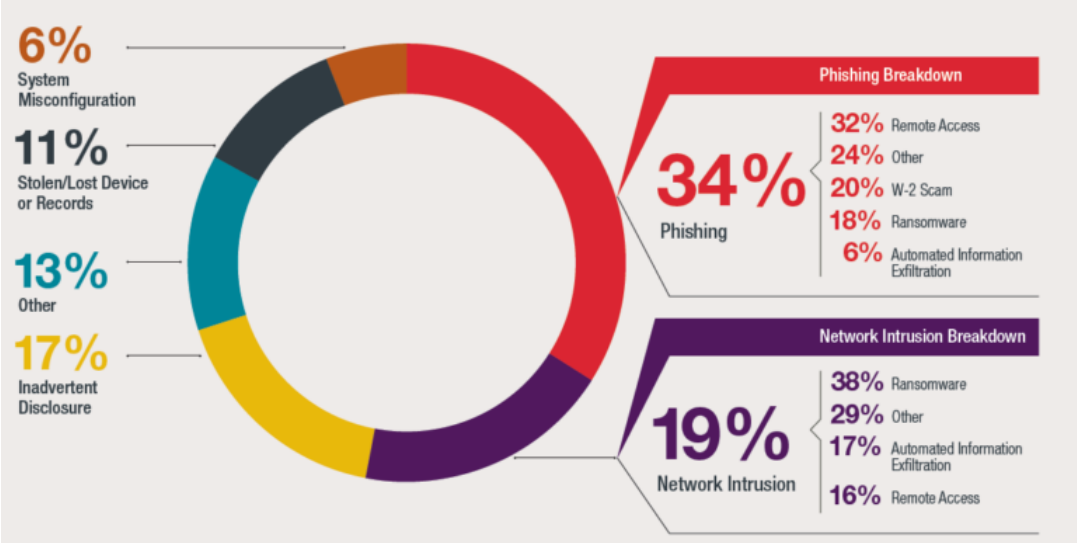

Figure 4 20: Main factor for data breach

Incidents Stats Platform

I have created all the graphs in the tool called Tableau. Tableau is used for the easy representation of the data. With the help of tableau, visualizing the graphs and data easily and it becomes easy to understand. SO to put all the things together I have created a dashboard where all the important data is represented in the form of statistics which makes the reader easy to understand.

Here are the public link and live dashboard for the tableau dashboard where the graphs can interact.

https://public.tableau.com/views/USDataBreachIncidents/Dashboard1?:language=en&:display_count=y&publish=yes&:origin=viz_share_link

- US Incidents

  The dashboard below shows the important factors which lead to data breach the factors as several incidents by each state, by type of breach, location, and year.

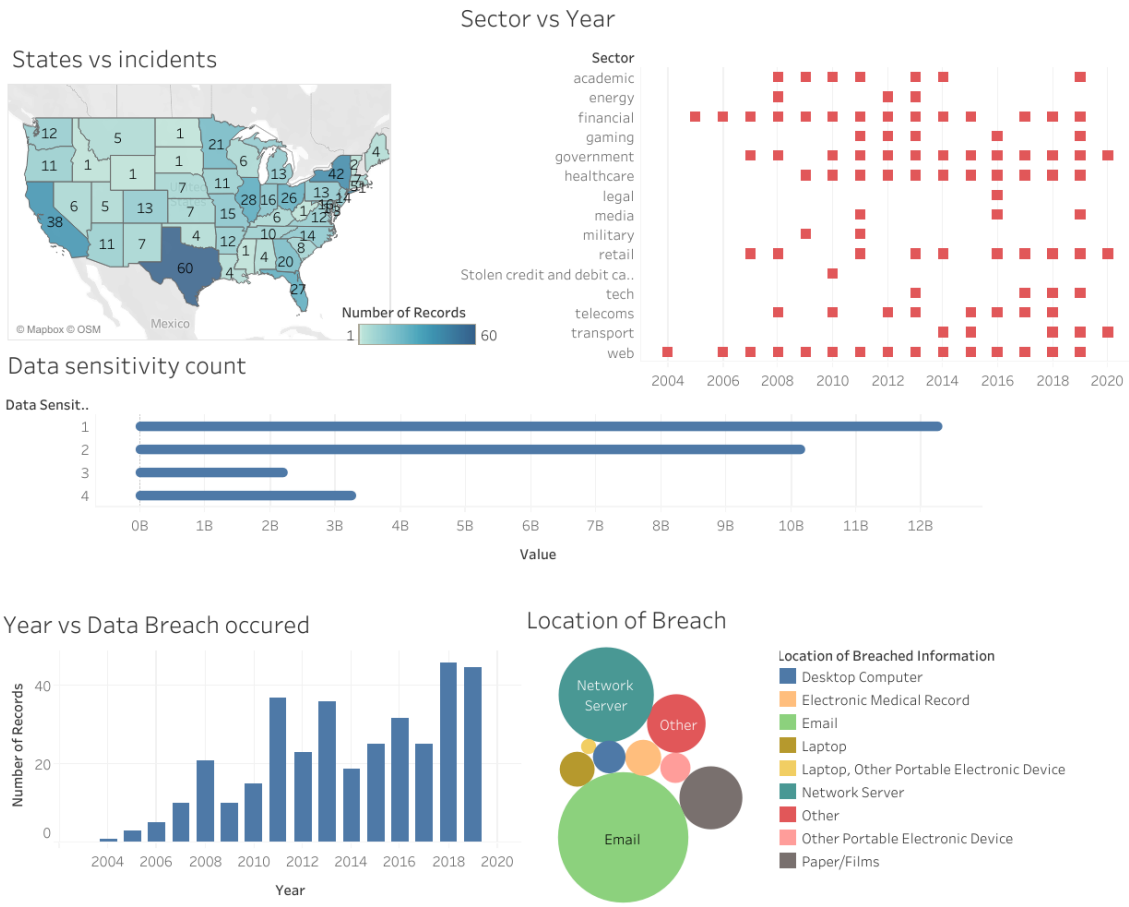

Figure 4 21:US Incidents at a Glance

CHAPTER FIVE

RECOMMENDATION

Bit by bit directions to Prevent Security Breaches Network security danger are consistent and authentic. By essentially using the web, people are persistently being shelled by various kinds of web risks. A wide scope of web threats applies various kinds of malware and distortion, interfaces in email or messaging, or malware associations that approach the Web. Focusing more on the Ultimate Guide to Network Security Threats With all the various sorts of framework security risks, it is possible to prevent security infiltrations by looking at the four mechanical assemblies which are used by organizations presently.

Types of Tools

SIEM

There is a need for SIEM to help log security events for affiliation. This is the mainline of shield to prevent security breaks. It is recommended for most of the industry to use SIEM as the main factor to keep a log of events. The affiliation likely has firewalls, IDS/IPS, and AV plans presented that scan for harmful attacks at various levels inside the IT organization, from the outskirt to endpoints. In any case, an extensive part of these game plans is not set up to perceive zero-day attacks and advanced serious dangers. Help prevent security breaks by adding SIEM development to the weapons store.

What is SIEM?

Security Information and Event Management (SIEM) – A SIEM stage halfway assembles data from various contraptions on the framework, including the present security machines. Through an impelled relationship engine, it can proactively recognize security events for no situation perceived by free security development. A SIEM structure fuses logging limits on security events for adventures and is used to look at or possibly report on the logged areas. The examination capacities of SIEM systems can recognize ambushes not found through various techniques and can organize the reconfiguration of other endeavor security controls to organization security. A part of the top SIEM things is expecting an attack is still in progress and can even stop data breaches.

There are many reasons to consider Managed SIEM including:

- Finding and keeping up experienced SIEM/SOC Security Analysts which isn't easy (and expensive).
- It could be built, yet it will take any more extended than redistributing to a specialist security organization provider.
- Getting everything from an MSSP exactly at a limited quantity of what could spend inside Scalable and Flexible.
- More notable Threat Intelligence has been doing this for quite a while and this has shown plenty of things.

Endpoint Detection and Response (EDR)

EDR hinders security breaks with endpoint disclosure and response. The gathering utilizes man-made mental aptitude that will help stop impelled risks and malware at the weakest point which is the endpoint. Antivirus isn't adequate to make sure about the endpoints. The concealed advancement for Cybriant's EDR organization is the fundamental development that stops by 99% of forefront threats and malware before they can execute to cause hurt. It takes out the prerequisite for legacy antivirus programming and prevent the data breach, whitelisting courses of action, and host-based interference ID and avoid systems. The use of "desire first" progression which is to stop assaults before it can cause more damage, instead of permitting ambushes to occur. By decreasing the measure of endpoint security things sent on the endpoint, clients increase operational efficiencies by not managing engravings, methods, or courses of action of extra insurance. EDR can help take out legacy endpoint security development that is not suitable against the current peril issues, along these lines improving cost speculation reserves and the board overhead. The development was attempted by HIPAA security assessors and it was better than any antivirus software which prevent malware attacks.

Managed Endpoint Detection and Response Benefits

At the point when re-appropriate the administration of the Endpoint Detection and Response (EDR), security experts suggest that:

- Perform underlying driver investigation for any blocked danger or some other ancient rarity regarded significant found on an endpoint.
- Proactively look endpoints for indications of dangers usually alluded to as danger chasing.
- Take unequivocal activity when a security episode, or possible occurrence, is distinguished.

Patch Management

The number of progressing cybersecurity breaks that have some answers concerning the news has been achieved by known vulnerabilities that ought to be fixed. According to a progressing Poneman study, "To prevent data breaks, security bunches need to fix even more quickly," the assessment says. "Regardless, the examination shows that they are being held somewhere near manual methods and withdrew structures that deal with their ability to fix helpfully." Patching the officials is a fundamental methodology in an IT organization which can help to prevent security breaks, this can have the best impact. The best way to deal with ensure suitable fix the board is to redistribute to an association like Cybriant and use automation. Responsive Patch Management course of action will channel the structures, check for missing and open patches against the comprehensive defenselessness database, download and send missing patches

and organization packs, and produce reports to feasibly manage the fix. The Responsive Patch Management course of action handles each piece of Windows, Mac, Linux, and other rare applications to fix the board. This consolidates sending patches faultlessly across work territories, PCs, servers, wandering contraptions, and virtual machines, from a singular interface. Responsive Patch Management game plan will invigorate the arrangement benchmark definitions to consolidate the new fixes, regularly separate to ensure that all endpoints remain consistent, perceive redesigns, and re-try the fix the official's methodology in like way.

Vulnerability Management

PCs and servers are not the only cause for data breaches. It is at present a marvelous mix of cutting edge enlisting stages and assets which address the bleeding-edge attack surface, including cloud, holders, web applications, and mobile phones. Proactively find certifiable asset characters (rather than IP addresses) over any propelled figuring condition and maintain a live point of view on the main points with the help of an official organization. Performing only a single shortcoming analysis each year or quarter places relationships at risk for not uncovering new vulnerabilities. The time between each clear is irregular. The attacker needs to deal with a framework. With constant shifting, the security experts have the detectable quality to assess where each advantage is secure or revealed. By using danger prioritization, the security experts have the solution to manage the settings. The enumerating will help with sorting out which exposures to fix first if using any means and apply the best possible remediation strategy. The

bleeding edge ambush surface has made a tremendous gap in an affiliation's ability to grasp its advanced presentation. Greater the gap, the more critical the risk of a business-influencing computerized event occurs. Standard Vulnerability Management isn't, now sufficient. Regulated Vulnerability Management widens vulnerable board by covering the extensiveness of the ambush surface (IT, Cloud, IoT/OT) and give a sign of information into the data (checking prioritization/examination/decision help). On the off chance that is set up to hinder security enters for the affiliation, consider PREtect. It provides layered assistance that offers all of the four things in a versatile and sensible computerized danger to the board organization.

Policy Recommendation

Keep Only What You Need

Whenever the user is done with the information or anything that the user does is that just delete or throw the thing or the information in the trash which might be harmful in some cases. So whenever the user wants to dump any information or any other thing which might reveal the personal information always burn or shred the material to save the people from identity theft.

Safeguard Data

Data is very important as it contains more valuable information that money. So to safeguard it the user can protect it with some kind fo password or key. Always keep the online account's password protected which is easy to remember and difficult to crack.

Control Computer Usage

Confine representative use of PCs to business use. Try not to allow workers to utilize document sharing distributed sites or programming applications, square access to improper sites, and avoid utilization of unapproved programming on organization PCs.

Secure All Computers

Install secret key insurance and break capacities for all PCs. User needs to train the workers to never leave the machine and the computers alone as it might help in the leaking of information. Confine working from home to organization possessed PCs. There should be a requirement of strong passwords which should be changed every quarter. Try not to store individual data on a PC associated with the Internet except if it is basic for directing business.

Never Unencrypted Data Transmission

So if the data is not encrypted it should not be transferred from an unsecured network unless the data is not so important. This incorporates information 'very still' and 'moving'. Additionally consider encoding email inside the

organization if individual data is transmitted. Prevent utilizing Wi-Fi systems; they may allow capture attempts of information.

Careful Usage of Portable Media

Nowadays there any many electronic items which have a memory inside it. This storage can be harmful because when they abandoned device the memory inside it doesn't get erased which leads to data identity theft. Use the pen drives, portable hard drives, and the thumb drive carefully and never connect it to the unknown USB slot. Permit just encoded information to be downloaded to versatile capacity gadgets.

Administrative recommendation

Educate/Train Employees

Set up a composed strategy about protection and information security and impart it to all representatives. Expect representatives to take care of records, log off their PCs, and lock their workplaces/file organizers by the day's end. Teach workers about what sorts of data are delicate or classified and what their duties are to secure that information.

Update Procedures

Try not to share the social security number as the primary ID just find an alternative that shows that it as an ID proof. If it is done that way, build up another ID framework right away.

Human Behavior Recommendations

Keep Security Software Up-To-Date

Always keep the electronic devices updated. Use firewalls, hostile to infection, and against spyware programming; update infection/spyware definitions every day. And do not update the device form an unknown source. Always download from the verified source.

Destroy Before Disposal

Always remember everything of the user has data inside it. It includes a paper to any electronic device. It is recommended that if the user dumps any old paperwork or any letters which the user doesn't need just use shredder so that the document doesn't give away any important information. Likewise, be aware of copiers, the same number of this output a report before duplicating. Change the settings to clear information after each utilization.

CHAPTER SIX

FUTURE WORK

System Design

In the Figure 6.1, This is generalized flow of how to filter email.

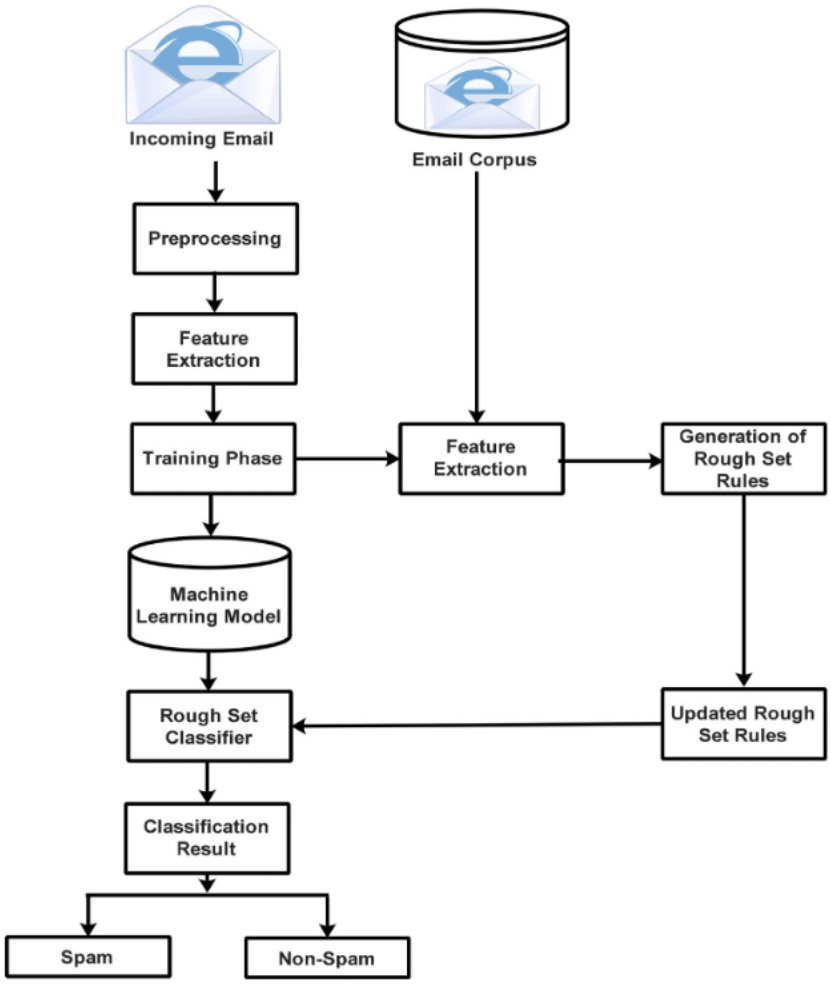

Figure 6 1: Email filtering system

Workflow for HKNI System (Hook and Eye System)

In this flow, there is the proper way the email will flow when it will be received from outside organizations to the receiver.

System Explanation

1. Firstly the email will be received to the HKNI system.
2. The email will be searched for links
3. If there is any link it will be changed for if it has text or graphics.
4. If there is no link in the email then it will be just checked for malicious words or graphics.
5. Then if there are any text, it can go to HKNI or if there are any graphics it will go to check for image extension.
6. If there are any graphics it will again go back to check image extension or else auto-check for exploit.
7. The one with malicious links or words will be discarded automatically and then manually checked.

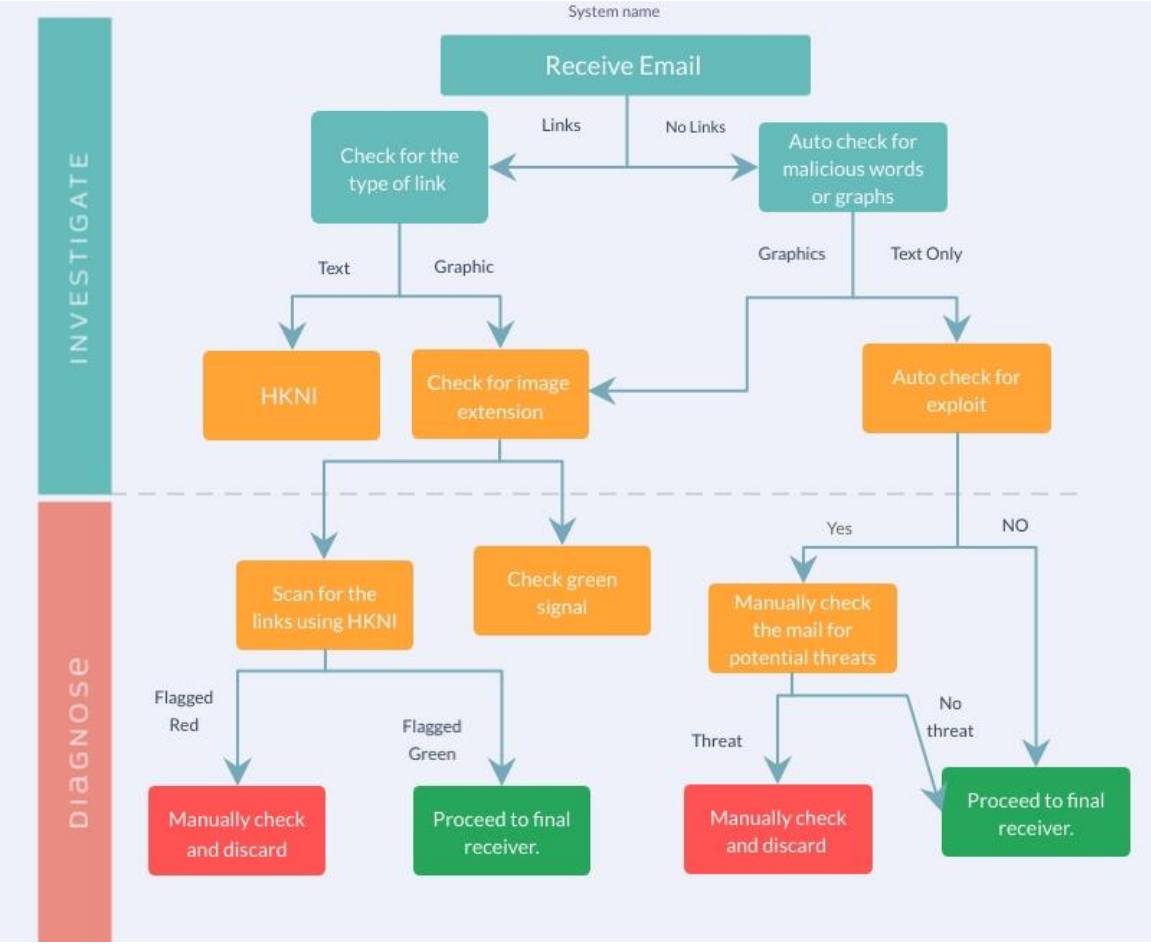

Figure 6 2: HKIN System

Website Design

In this chapter, the workflow of HKNI systems is discussed. For that, the basic requirement for the system to work is to develop a website. For this website, HTML and CSS scripting language is used. With the combination of both, the use of web scrappers and web crawlers is done to identify the high-risk words. And all this will be hosted on a private server to test and use of XAMPP servers and MySQL database to store data. This is the basic idea of how the website is going

to work and look like. Screenshots are shared which has mentioned below which will help the reader to understand easily.

Login Page

Firstly, the look and feel of the website are being discussed, but for now, this will be only an idea where some photos of the portal and how it looks will be shared. In this chapter, the website looks, and works are discussed. So, starting with the main page, the page will consist of login for admin. As the user won't be allowed to log in as this system is completely controlled by the cybersecurity team. There will be 2 access to the website. One with the admin and the other with another security team that would have limited access. The admin team will have full access. So the main purpose of this page is to have security for the login of this website so that any other person from the organization shouldn't be able to login to the system.

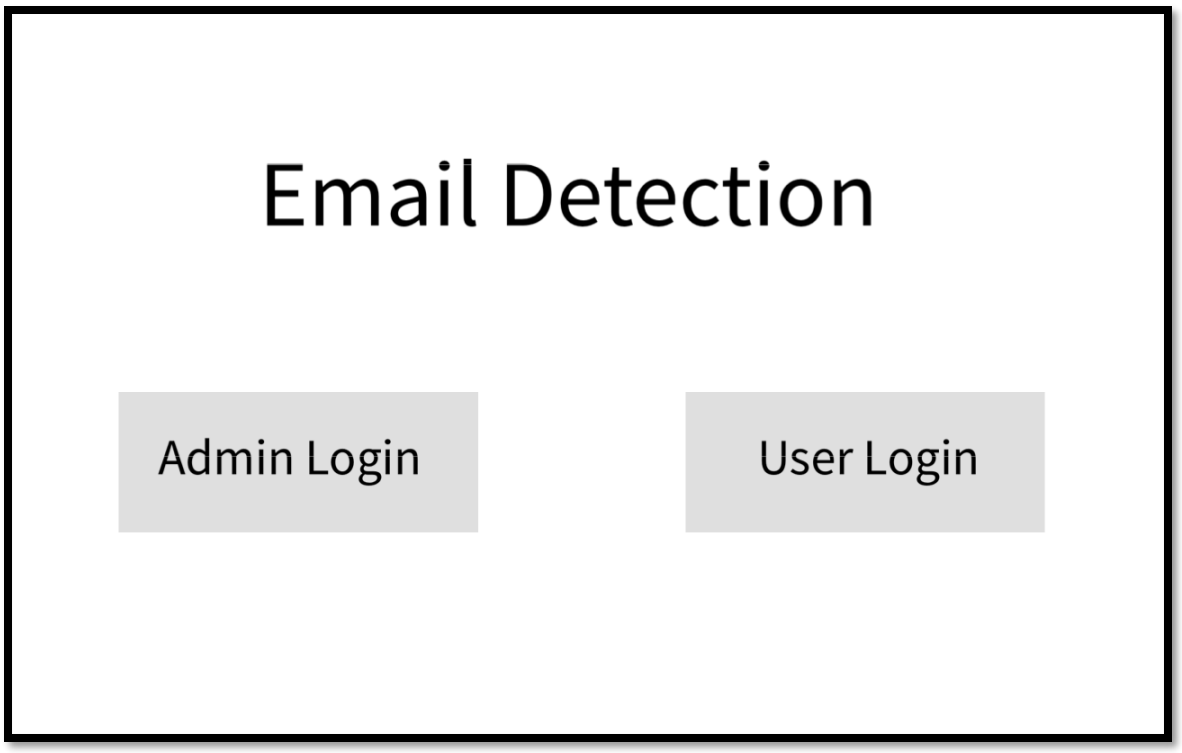

Figure 6 3: Login page

Credential Input Page

On this page, the admin can input the credentials for the account so that he can log in to his account. After a successful login, the page will be redirected to another page which will be the main interface. The login page will be a PHP code written in the sublime text editor with the help of a server created on the machine using XAMPP.

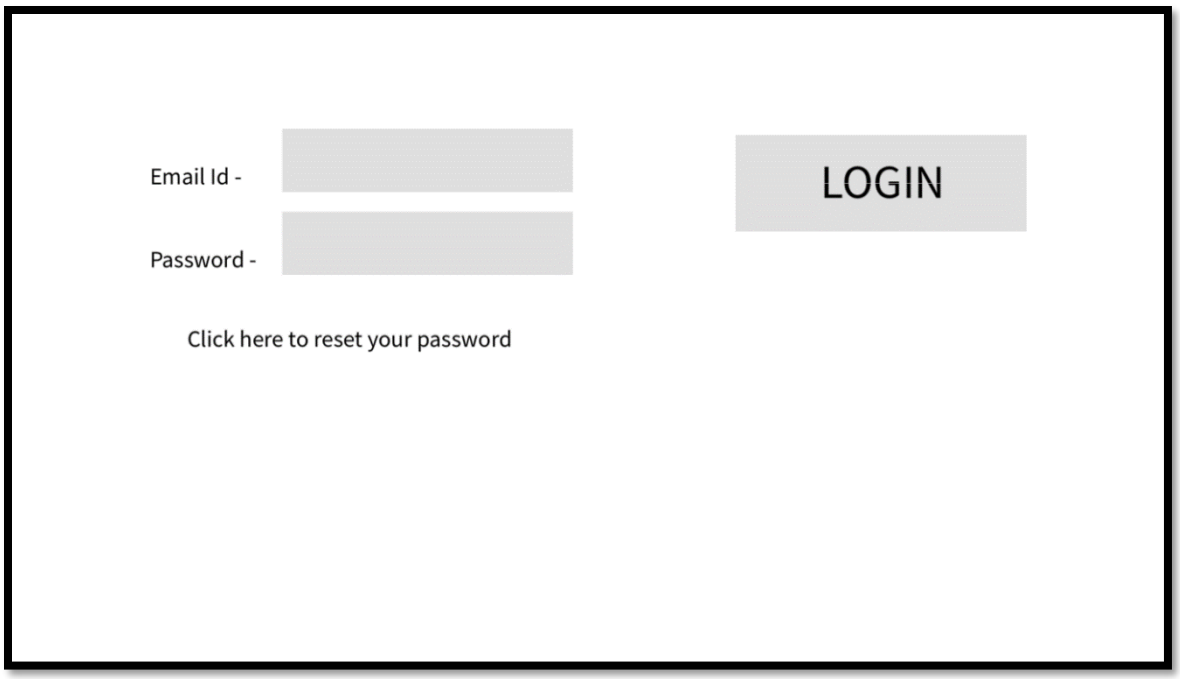

Figure 6 4: Credential Input page

Code for the login page

```php
<?php
   include("config.php");
   session_start();

   if($_SERVER["REQUEST_METHOD"] == "POST") {
      // username and password sent from form

      $myusername = mysqli_real_escape_string($db,$_POST['username']);
      $mypassword = mysqli_real_escape_string($db,$_POST['password']);

      $sql = "SELECT id FROM admin WHERE username = '$myusername' and passcode = '$mypassword'";
      $result = mysqli_query($db,$sql);
      $row = mysqli_fetch_array($result,MYSQLI_ASSOC);
      $active = $row['active'];

      $count = mysqli_num_rows($result);

      // If result matched $myusername and $mypassword, table row must be 1 row

      if($count == 1) {
         session_register("myusername");
         $_SESSION['login_user'] = $myusername;

         header("location: welcome.php");
      }else {
         $error = "Your Login Name or Password is invalid";
      }
   }
?>
<html>
```

Figure 6 5: Code for credential page

Main Interface

The interface consists of the main menu and the function in it. The main interface contains View User, Add to the blocklist, add words, view list, view blacklist, view feedback, logout. The admin will be able to see the list of all the users in the organizations and will be able to see all the emails coming and going to monitor all the movements going on. Admin will be able to do all sorts of adding, deleting, and updating the data and accounts on his side.

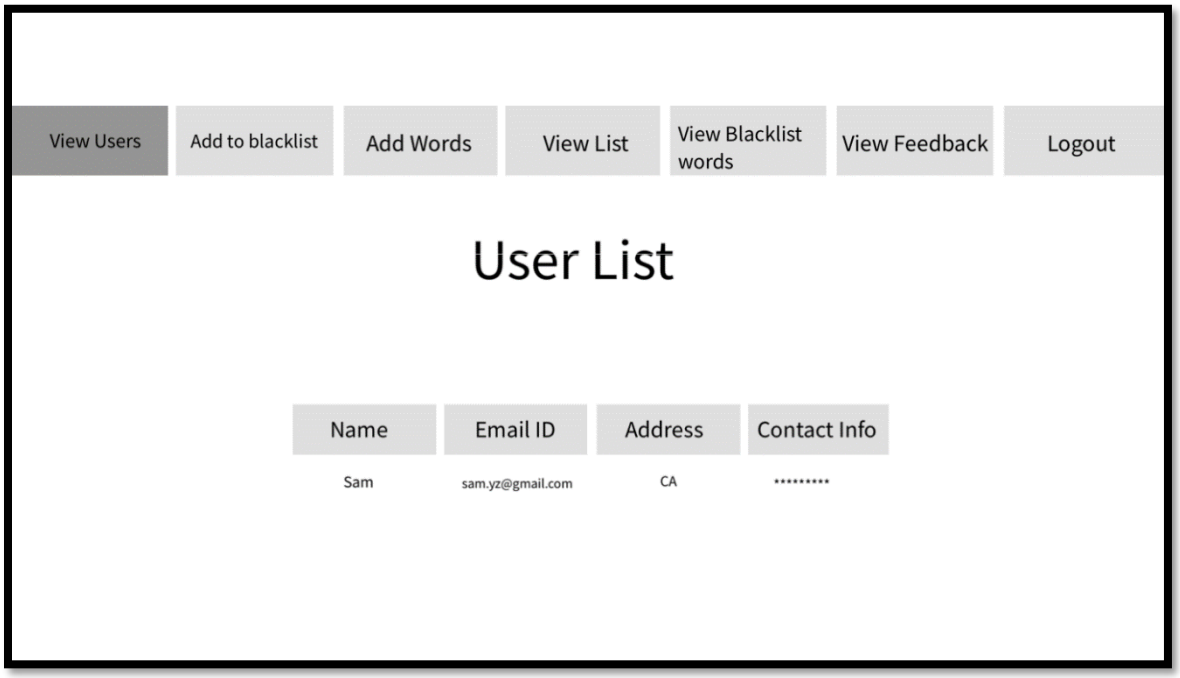

Figure 6 6: Main interface

Add to blacklist Interface

On this page, if any malicious links were found on the HKNI system it will get added directly to this box and will be added to the database. The web scrapper

will detect the malicious link and check it with the available links in the database. If there is not the HKNI system will add it to the database.

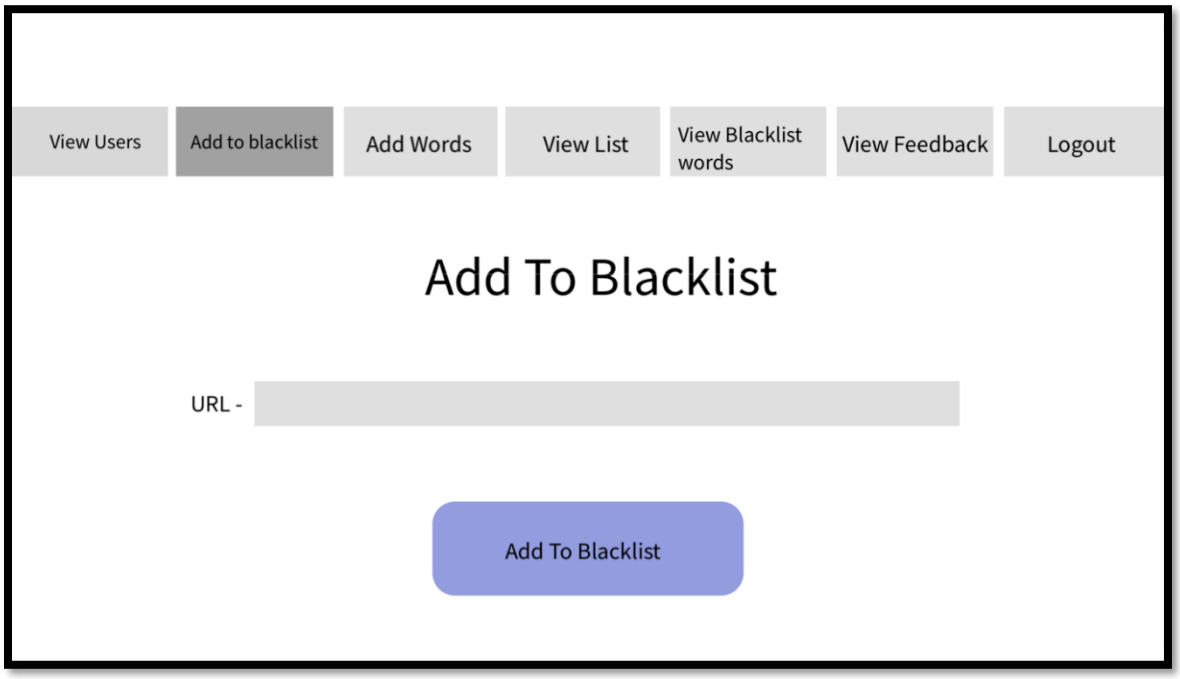

Figure 6 7: Add blacklist page

Add Keyword Interface

On this page, if any malicious words were found on the HKNI system it will get added directly to this box and will be added to the database. The web scrapper will detect malicious words and check them with the available words in the database. If there is not the HKNI system will add it to the database.

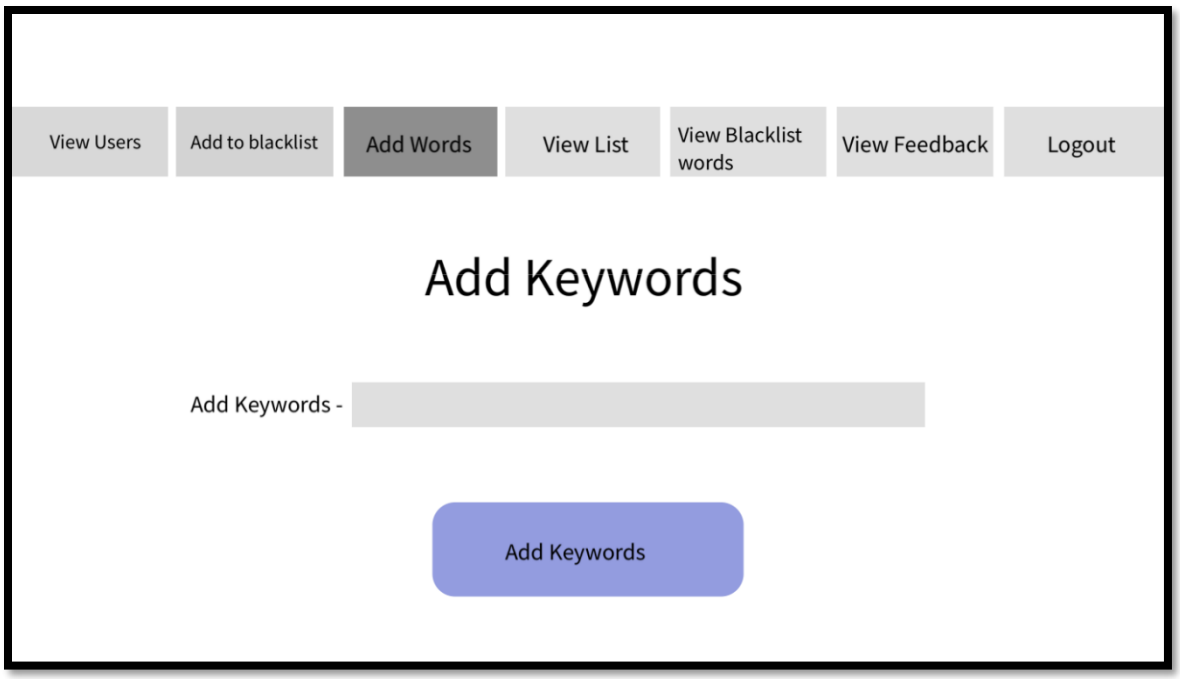

Figure 6 8: Add keyword interface

View Keyword List

On this page, it is viewed that all the words which are in the database and also manually add other malicious words id needed.

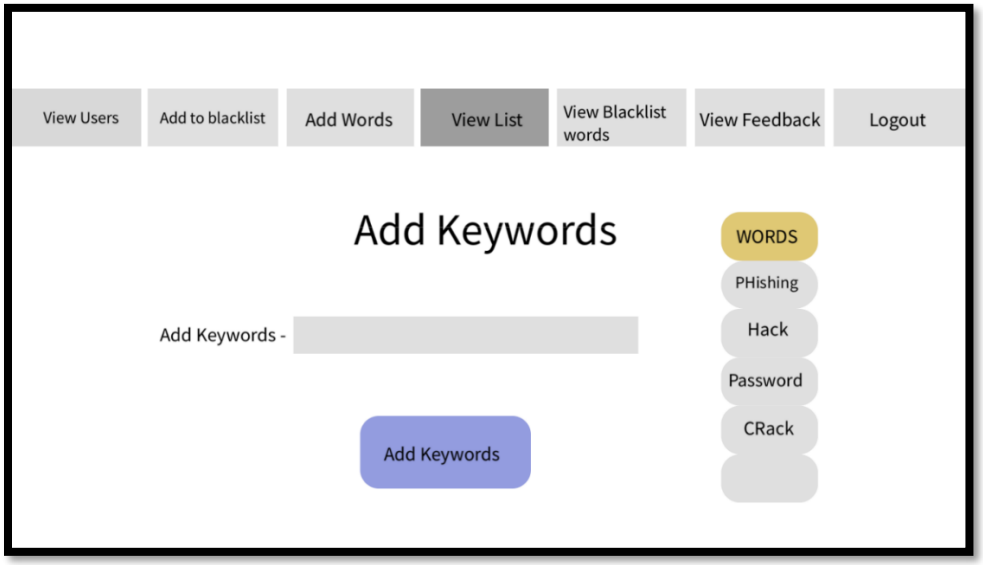

Figure 6 9: Keyword interface

View Blacklist link List

On this page, all the links which are in the database and also manually add other malicious links id needed.

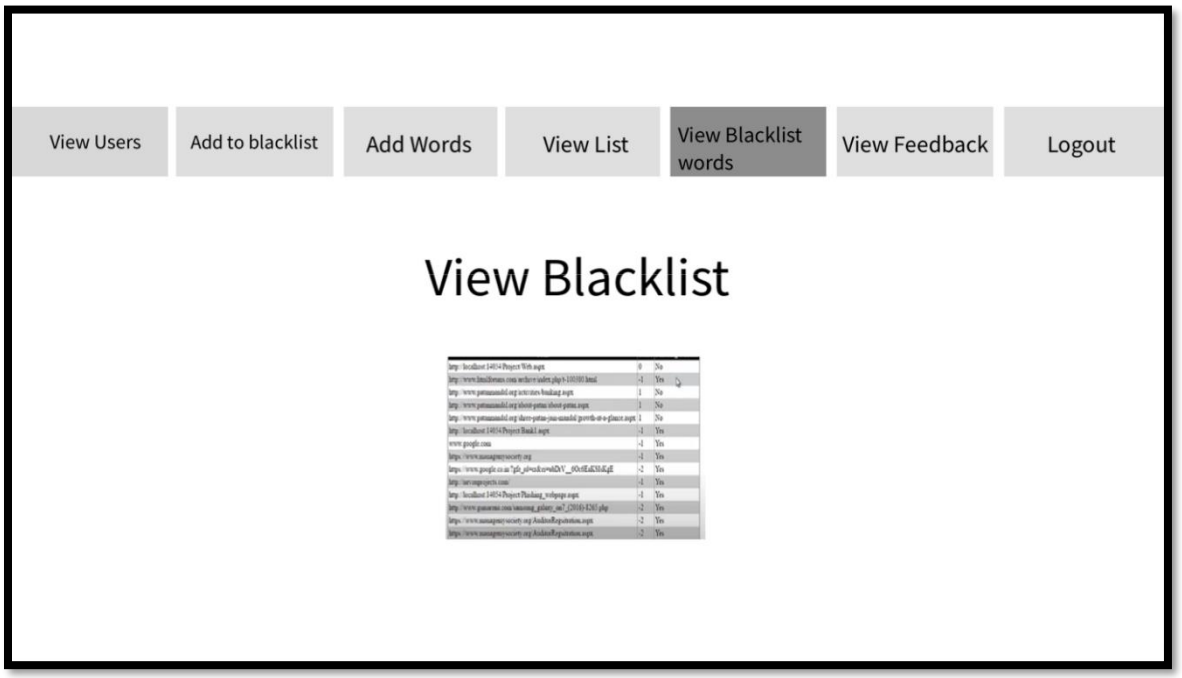

Figure 6 10: View blacklist page

Use case Diagram

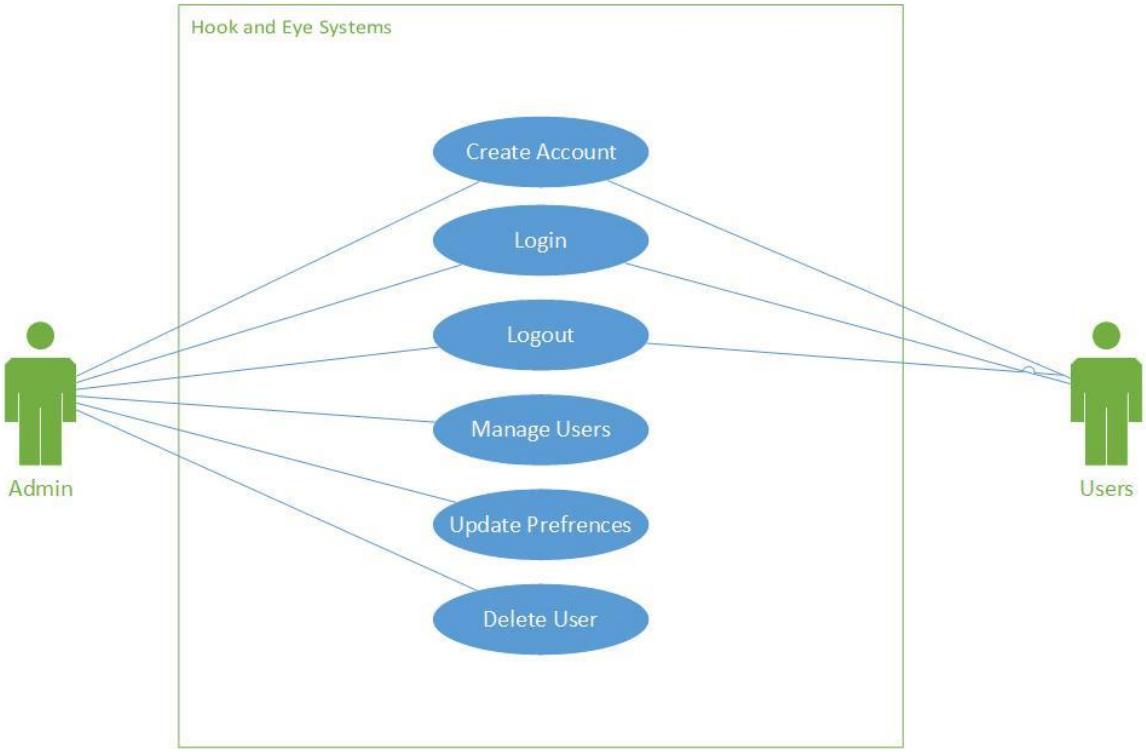

Figure 6 11:Use case

Figure 6.11 shows that admin has access to the system with total authorities and the other security engineers and analyst has minimum access. This system will give full access to the admin which would control the whole systems which include access to the keyword database and blacklisted link database.

CHAPTER SEVEN CONCLUSION

From all these analyses it is found out that the data breach incident reports are increasing day by day and in this COVID-19 crisis there is a substantial increase in the number of attacks on the identity of a person. From different types of analysis, it can be a conclusion that the data breach through phishing and emails is the most common type of attack which hackers use to get into the system. The most common methods to prevent these types of attacks are to spread general knowledge of how to handle the technology and how to keep everything updated. Some common practices are explained in the project above.

Form the analysis it has been concluded that misuse of email can cause a great threat and lead to disaster. As the people know there are several systems designed to prevent the phishing attacks, in this project there is the design of the HKNI system which uses a machine learning model and would help the organization to reduce the impact of the data breach by emails by manually and automatically scanning the emails and comparing it with the existing database with all the vulnerable links and words and also keeping it updated with the new words found in the emails. As this system is just the idea of how the organization can upgrade their security systems there is a web-based portal which can do the exact mechanism explained in the project. There are some drawbacks to the system which is the time required for one email to undergo the process is still unknown which can affect many factors.

REFERENCES

Cipher. 2020. *5 Effective Ways To Prevent Data Breaches - Cipher*. [online] Available at: <https://cipher.com/blog/5-effective-ways-to-prevent-data-breaches/> [Accessed 27 June 2020].

Compliance, S. and Security, H., 2020. *How To Build An Effective Data Classification Policy For Better Information Security*. [online] Blog.netwrix.com. Available at: <https://blog.netwrix.com/2018/05/31/how-to-build-an-effective-data-classification-policy-for-better-information-security/> [Accessed 27 June 2020].

Cybriant. 2020. *4 Necessary Tools To Prevent Security Breaches | Cybriant*. [online] Available at: <https://cybriant.com/prevent-security-breaches/> [Accessed 27 June 2020].

Irwin, L., 2020. *How Do Data Breaches Happen? Understanding Your Organisation'S Biggest Threats - IT Governance UK Blog*. [online] IT Governance UK Blog. Available at: <https://www.itgovernance.co.uk/blog/understanding-the-different-types-of-data-breaches> [Accessed 27 June 2020].

Safecomputing.umich.edu. 2020. *Examples Of Sensitive Data By Classification Level / Safecomputing.Umich.Edu*. [online] Available at: <https://safecomputing.umich.edu/protect-the-u/safely-use-sensitive-data/examples-by-level> [Accessed 27 June 2020].

The AME Group. 2020. *Data Security Breach: 5 Consequences For Your Business*. [online] Available at: <https://www.theamegroup.com/security-breach/> [Accessed 27 June 2020].

Upguard.com. 2020. *What Is Sensitive Data?*. [online] Available at: <https://www.upguard.com/blog/sensitive-data> [Accessed 27 June 2020].

Us.norton.com. 2020. *What Is A Data Breach?*. [online] Available at: <https://us.norton.com/internetsecurity-privacy-data-breaches-what-you-need-to-know.html> [Accessed 27 June 2020].

Campbell, K., L. Gordon, L. Loeb, & L. Zhou (2003) "The Economic Cost of Publicly Announced Information Security Breaches: Empirical Evidence from the Stock Market," 11(3) J. of Computer Security 431.

Acquisti, A., A. Friedman, & R. Telang (2006) "Is There a Cost to Privacy Breaches? An Event Study," presented at the Fifth Workshop on the Economics of Information Security, Robinson College, University of Cambridge, Cambridge, UK

Kannan, K., J. Rees, & S. Sridhar (2007) “Market Reactions to Information Security Breach Announcements,” 12(1) International J. of Electronic Commerce 69.

Gordon, L. A., M. Loeb, & L. Zhou (2011) “The Impact of Information Security Breaches: Has There Been a Downward Shift in Costs?” 19(1) J. of Computer Security 33

Belsis, P., Kokolakis, S. and Kiountouzis, E. (2005), “Information systems security from a knowledge management perspective”, Information Management & Computer Security, Vol. 13 No. 3, pp. 189-202.

CERT (2008), Governing for Enterprise Security, Computer Emergency Readiness Team, available at: http://cert.org/governance (accessed 12 August 2010).

Pike, G. (2009), “Congress debates data breach legislation”, Information Today, Vol. 26 No. 11, pp. 17-19.

M.T. Khorshed, A.S. Ali, S.A. Wasimi, A survey on gaps, threat remediation challenges and some thoughts for proactive attack detection in cloud computing, Future Gener. Comput. Syst. 28 (6) (2012) 833–851.

G. Gonzalez-Granadillo, S. Dubus, A. Motzek, J. Garcia-Alfaro, E. Alvarez, M. Merialdo, S. Papillon, H. Debar, Dynamic risk management response system to handle cyber threats, Future Gener. Comput. Syst. 83 (2018) 535–552.

Shu, X., Tian, K., Ciambrone, A., Yao, D. (2017). Breaking the Target: An Analysis of Target Data Breach and Lessons Learned. CoRR,abs/1701.04940
Smith, T. T. (2016). Examining Data Privacy Breaches in Healthcare.

When we discuss incidents occurring on NSSs, are we using commonly defined terms?, "Frequently Asked Questions on Incidents and Spills", National Archives Information Security Oversight Office