# ALEKSI
# Data Leakage Prevention system

Project ID 046

Project Proposal

Proposal documentation submitted in partial fulfillment of the requirement for the Degree of Bachelor of Science Special (honors)
In Information Technology

Bachelor of Science Hons. In Information Technology
(Specialization in Cyber Security)

Department of Information Systems Engineering

Sri Lanka Institute of Information Technology
Sri Lanka

March 2017

# ALEKSI
# Data Loss Prevention system

Project ID 046

Project Proposal Report

Authors:


| Student ID | Name | Signature |
|---|---|---|
| IT14030918 | R.A.C.D Ranathunga | |
| IT15084064 | A.R.Jayaweera | |
| IT15017598 | S.P.D Kaveendra | |
| IT15038388 | Weerasuriya.A.A.H.W | |


Supervisor

…….

Amila Nuwan Senarathna

Co-Supervisor

…….

Kavinga Yapa Abeywardhana


Bachelor of Science Information Technology Specialization in Cyber Security

Department of Information Systems Engineering

## Declaration of the Candidate and Supervisor

We declare that this is our own work and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

| Student ID | Name | Signature |
|---|---|---|
| IT14030918 | R.A.C.D Ranathunga | |
| IT15084064 | A.R.Jayaweera | |
| IT15017598 | S.P.D Kaveendra | |
| IT15038388 | Weerasuriya.A.A.H.W | |

The above candidates are carrying out research for the undergraduate Dissertation under my supervision.

Signature of the supervisor: Date

## Abstract

As the value of data and information growing rapidly with the technology, attacks
targeting information system has increased. In order to overcome this threat and protect data from data leakages, organizations use data leakage prevention solutions to control access to sensitive data and monitor the data flow.
Most of the current DLP solutions are using anomaly based and signature-based approaches
But the problem with these approaches is that they generate a lot of false positive
Results, these solutions are extremely high capital demanding that medium to small size organizations are unable afford these DLP solutions. To overcome these issues in current DLP solution we propose a solution based on machine learning to identify data leakages. Through this approach
we aim to improve the accuracy of the detection as well as to provide a specifically trained DLP solution that is uniquely adapted to the
organization, which is able to perform Email analysis, cloud upload and data syncing analysis, Web upload analysis and removable storage media analysis.

# Table of Contents


| | |
|---|---|
| **Declaration of the Candidate and Supervisor** | III |
| **Abstract** | IV |
| **Table of Contents** | V |
| **List of figures** | VI |
| **List of tables** | VI |
| **List of Appendices** | VI |
| **Introduction** | 1 |
| **Objectives** | 7 |
| **Methodology** | 8 |
| **Description of Personal and Facilities** | 18 |
| **Budget and budget Justification** | 19 |
| **Conclusion and Recommendations** | 19 |
| **References** | 20 |
| **Appendix A: Gannet Chart for ALEKSI System** | 22 |

## List of figures

Figure 1. Causes of Security Breaches
Figure 2. Overall structure Diagram
Figure 3. Iterative Waterfall Model
Figure 4. Email analysis
Figure 5. The methodology in which web post detection system operate.

## List of tables

Table 1. Estimated average Revenue Impact of a Leak
Table 2. Personal and Facilities
Table 3. Budget

## List of Appendices

Appendix A: Gannet Chart for ALEKSI System

# 1.Introduction

## 1.1. Background and Literature Survey

Unauthorized parties gain access to Sensitive and protected data can be defined as Data leakages or Data breaches and attempt to prevent such incidents from happening can be called as data leakage prevention(DLP).

With the advancement in technology many government and private sector organization have moved into electronic methods to store and handle data to increase the efficiency and it has increased the value of data and information in the modern world. It also increases the risk of damage and the cost of data leakages. According to the survey done by Ponemon Institute over 25% of security breaches are caused by Negligent Employees and Negligent Third Parties.

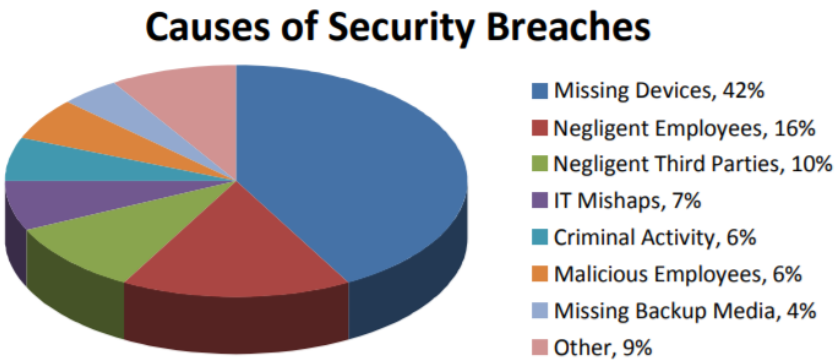


Fig 1. Causes of Security Breaches [1]

Data leakages always have negative effect on organizations and individuals. It's a huge threat to enterprise world because there are very sensitive data like trade secrets, financial data and other highly sensitive data alone with personal information. Some information like credit card, health information and customer names/ addresses leakages can breach the trust between customer and the vendors. Organizations have to face many severe effects like loss of customers, loss of organization reputation, fines due to violation of customer privacy. As shown in the table [Tab 1] below, data leakage incident can affect to the organization revenue also. According to Forrester research institute, organization can lose up to 20% customers due to a data leak.

| | Repeat Customers | New Customers | Total |
|---|---|---|---|
| Total annual revenue | $800 million | $200 million | $1 billion |
| Lost business as a percentage of revenues | 10% | 20% | 12% |
| Lost business in dollars | $80 million | $40 million | $120 million |

Table 1 : Estimated average Revenue Impact of a Leak [1]

One of the main issue with data leakages is that data leakages happen very often but many of them remain unnoticed if the data isn't published to the public. So that detecting sensitive data is very important. In order to detect sensitive data, there should be classification of data and it's important to identify the state of data [2].

There are three steps of data states. First state is data at rest (DaR) which is used to refer data stored in storage medium. It can be store on a hard disk, backup tapes or mobile device. These data are stable and there's no active transmission through the network.

Data in use(DiU) is the second state of data and it's also stored in some kind of storage media but actively interacting with one or more application. These data are being updated, deleted or processed.

The last state is data in motion(DiM) which is used to refer data that actively transmitted through the network. The data transmitting through the network as well as actively interacting with the system are constantly in the state of updating and in motion. Data loss prevention(DLP) systems mainly focus on data in this state. It's important to monitor data in motion to avoid data transferring to unauthorized parties.

There are many methodologies to prevent data breaches. Extrusion prevention system or exfiltration prevention is the method of preventing data leaks by monitoring the outbound traffic of the network and avoid unauthorized packets from moving outside the network.

Content monitoring and filtering (CMF) and Content monitoring and protection(CMP) is another method that used to prevent data breaches by monitoring the data in the organization and avoid malicious data transferring within the organization. There are other technologies like outbound content compliance and information protection and control which depends on organization policies. [3] There are many solutions to prevent data leakages and some of them are even hybrid solutions. Current data loss prevention solutions based on keyword search, sensitive data tagging, Digital fingerprints and machine learning.

EMC-RSA, McAfee, Symantec and Verdasys are some of the enterprise level DLP solutions. [4]

RSA Data Loss Prevention Suite is a DLP solution released by EMC Corporation which is operated in client endpoint. It performs scans on Microsoft SharePoint and monitors file shares and data repositories. Detecting sensitive data leakage process is completely based on keyword search and digital

fingerprints. But there are limitations in this product. RSA Data Loss Prevention Suite designed for only high-level protocols and It's mainly based on windows platform and It's not working on Linux platform. Also, there's no cloud and mobile infrastructure support in this product.

McAfee Data Loss Prevention is another popular DLP solution which performs on Data in motion to detect sensitive data leakages by monitoring the network passively and it's also based on McAfee Epolicy Orchestrator. But it's limited to email and web monitoring and supports for only 4000 ICAP connections and 30 concurrent SMTP connections. Basic mechanisms behind the McAfee Data Loss Prevention is keyword-based search, digital fingerprinting and data tagging. But there's no Linux and cloud infrastructure in this product.

Symantec Data Loss Prevention can operate in multiple platforms like windows and Red hat Linux. Symantec Data Loss Prevention for Web and Symantec Data Loss Prevention for email are two main products in Symantec data loss prevention. Symantec DLP available for mobile devices, social medias and cloud infrastructure but limited to specific vendors. It uses Support vector machine algorithms to build statistical models with sample documents. But this solution is limited to specific service providers, certain software vendors and file types. [5]

Verdasys Digital Guardian(DG) Data loss prevention system is basically based on keyword search, regular expressions search and Bayesian analysis. Specialty with this product is It have ability to work with unstructured data and multiple operating systems. It uses context-based data monitoring. Verdasys Digital Guardian agents detect special characteristics/context of an entity to categorize data.

Most of the solution based on keyword, regular expression search and data tagging with digital fingerprints. Number of false-positive alert rate is one of the main drawbacks in anomaly and signature based DLP solutions. Even though there are some solutions that are partly based on machine learning to overcome this issue, their machine learning application is limited to few features of the overall product. Most of these solutions are also limited to large scale organizations due to the high cost.

Solution that we suggested to overcome these issues is a product that is totally based on machine learning to detect data breaches that occur through emails, cloud uploads, Web uploads and removable media. By using a solution based on machine learning, number of false-positives can be reduced compared to current solutions and data loss prevention solution can be trained specifically to an organization. [6]

### 1.2. Research Gap

Most of the existing solutions are based on whitelisting or blacklisting models.
There's no full machine learning implemented data loss prevention system in the existing solutions. So that these existing Data loss prevention systems aren't specific to organizations.
These products detect data breaches and transferring sensitive files by analyzing keywords / regular expression search and data tagging / digital foot printing mechanisms. Most of existing DLP solutions support only few operating systems and few software platforms.
Most of the Data loss prevention solution follow mechanism like watermarking and adding embedded code to the original data to analyses image and multimedia files which involves modification of the original data.

### 1.3. Research Problem

All most all the endpoint data leakage detection systems currently available relies on the keywords-based searches and anomaly detection to detect the data leakages which results in large number of false positive alerts in compared to the number of true positive alerts generated [7] [8]. Large number of false positive alerts will increase the overall alerts count, which means a human analyst would have to spend a considerable amount of time investigating these alerts, and major proportion of time would be spent on the false positive alerts due to the overwhelming number of false positive alerts. This is very counterproductive, and that time could be used to perform deep dive investigations on true positive alerts if the time spent on false positive alerts was minimized. Amount of time spent on false positive alert investigation may impact the investigation process carried out by the analyst and may lead to scenarios where the analyst would discard a true positive alert as a false positive without doing a deep investigation due to the inadequate time he or she has to spend on investigating a single alert. These types of scenarios will cause undetected data leakages which may ultimately lead to situations where the organizations will lose their competitive advantages, reputational damages and law suits.

Minimizing the counter productiveness that occurs through large number of false positive alerts is a major concern for the traditional data leakage detection systems, and they have failed to overcome the challenge so far.

### 1.4. Solution

To overcome the challenge faced by the traditional data leakage detection systems, we propose Aleksi, a machine learning based data leakage detection system which is more capable of detecting data leakages with much higher degree of accuracy. That would produce less number of false positive alerts compared to the traditional solutions.

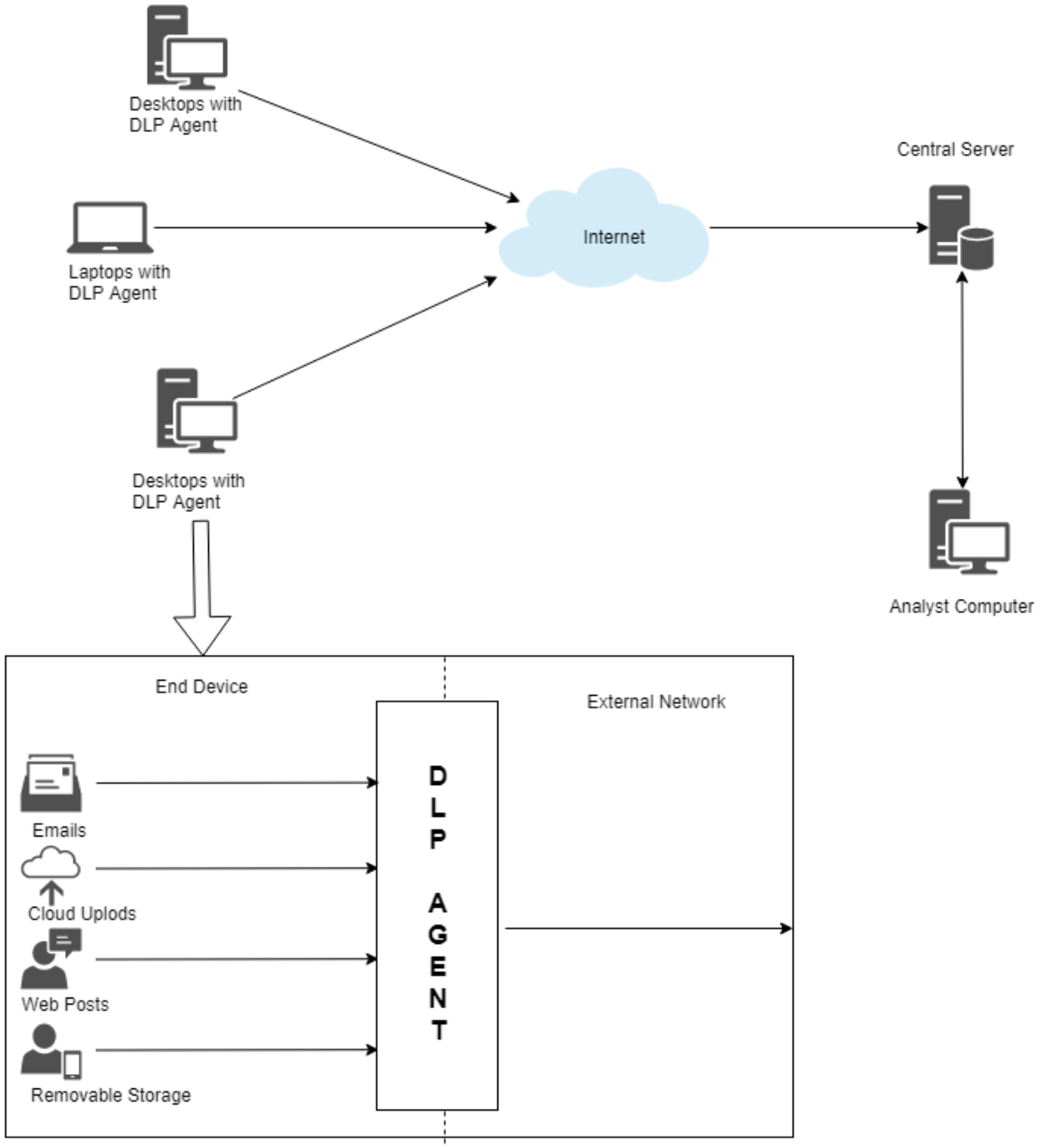


Figure 2: Overall structure Diagram

Keyword based data leakage detection is inefficient [8], a solution can be found through machine learning where data leakage patterns unique to an organization can be identified through a machine learning algorithm that is trained on a data set that is taken from the organization. Algorithm is trained on preselected data sets using supervised learning. This will make the algorithm more uniquely adapted to the organization that it is deployed on and will increase the accuracy of detection of data leakages.

The Algorithm will reside in the Aleksi agent and will run in the background of all the endpoints where the Aleksi agent is installed. Aleksi agent will act as an interface between endpoints' communications and the outside network and monitor the outgoing traffic of communication channels like Emails, Web posts, Cloud storages and Removable devices for any indication of data leakage. If it detects a possible data leakage pattern in the ongoing data an alert is triggered and sent to central server along with the evidence for the suspected data leakage. Analysts will be notified by the central server, from there analysts will be able to access the central server and start the investigation on the possible data leakage alert.

# 2. Objectives

## 2.1. Main Objectives

The main objective of Aleksi is to identify a data leakage with much higher accuracy than the traditional Data leakage systems and solutions.

A Data Leakage detection system should be able to identify any form of electronic data leakage that occurs through the network. Aleksi concentrates on 4 main vectors of data leakage that occur in the organizations today as

1. Detecting data leakages occur through Email.
2. Detecting data leakages occur through Web Posts.
3. Detecting data leakages occur through Cloud uploads.
4. Detecting data leakages occur through Removable Media.

Correctly identifying the instances of data leakage and capturing the evidence of the leak for investigations and minimizing the workload of the the analyst are the main objective of the Aleksi.

## 2.2. Specific Objectives

Almost all DLP Solution in the present are highly expensive and requires lot of capital to implement [9], lot of small to medium scale companies are unable spend on data leakage prevention solution or are unable to fully cover the entire organization from data leakages due to the high cost of the products available, our objective is to produce a solution which is economically feasible even for a small scale company so it would be able to protect its data from unnecessary leakages.

Another objective of the Aleksi is to provide unique product which is more flexible and adaptable to the nature of the organization's data and information. Aleksi addresses a critical weakness in traditional solutions which provide generic solutions independent of the nature of the data an organization may use. Aleksi achieves this through training on datasets that are taken from the organization it is being deployed.

# 3. Methodology

### 3.1. Overall procedure of the project

Aleksi system is designs to implement as a client-server architecture application. The Aleksi system is able detect organization's sensitive data when data is moving through Emails, Removable media, Clouds uploads and Web posts. System use set of pre-defined classification structure for analyzing process. In order to get the system work efficiently, it needs to be configured accordingly to the organization's data classification. After the proper configuration, system needs be trained with sample set of data to get high accurate results as output. After the training process system will classify sensitive data if they get detected during the analyzing process and send alerts to the main server. From the main server, analyst can take action according to the alerts he/she received. System will monitor data which transferred through removable devices, sent via Emails, uploaded to cloud storages and posted to web. This approach will help all the main four sub-system to use these modules for their analyzing process. System will have two main modules for its analyzing process.
They are;

1. Text analyzer.

    This module is designed to analyze text-based data using Naive Bayes machine-learning algorithm. Using Naïve Bayes algorithm, System will detect sensitive data and It will classify them into proper categories.

2.Image analyzer.

    This module will be responsible for identifying and classifying images that are detected through the Aleksi agent. The image classifier is trained using transfer learning [10] . It is training an already trained algorithm which has higher accuracy to identify and classify new image types. The module will be able to classify Personally identifiable information(PII), Protected Health Information(PHI), Source code images, Internal and client related confidential images, blueprints or design documents. By using transfer learning, System will be able to achieve higher accuracy with a small set of data [11]. TensorFlow will be used to classify the images as it is a well-known and a successful algorithm which has been trained already [12].

3.File Analyzer.

    Aleksi system capable of analyzing most widely used common files with following extensions. ".doc", ".docx", "xlsx", "xls", "ppt", "pptx" and PDF files.

One of most important factor for the success of the project is the Software Development Life Cycle (SDLC) methodology it follows. We choose to follow Aleksi system's SDLC based on Iterative waterfall methodology. The main reason for selecting this particular model is that, it provide ability make changes to the previous phases (developing stages) if needed.

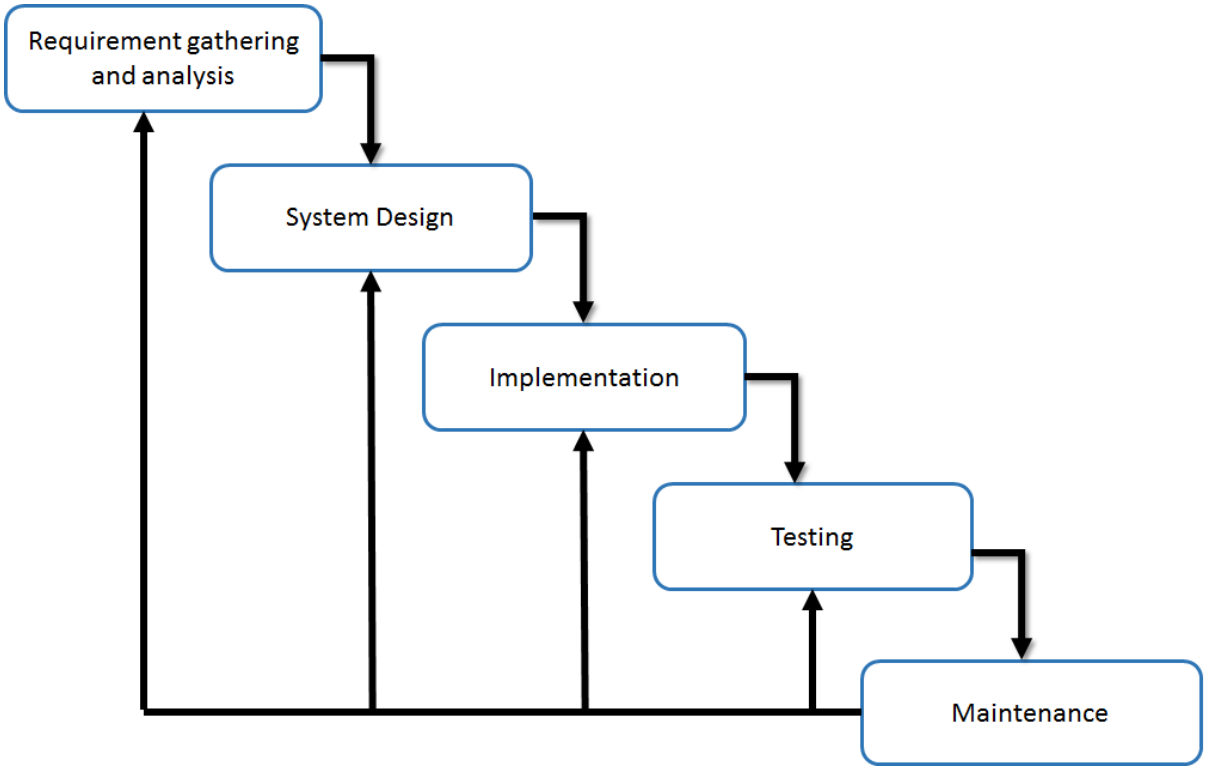


Figure 3 - Iterative Waterfall Model

**1. Requirement gathering and analysis**

Requirement gathering, and analysis phase is the most important phase in the SLDC since it is the phase where we can review limitations and problems of existing systems. During the requirement gathering phase we can identify all the requirements for the system and drawbacks which are existing in current solutions. According to our research on current solutions, there is no fully functional data loss prevention system that uses machine-learning technology. During this phase we need to identify functional and non-functional requirements. Following activities are planned to carry out within this phase to make sure that all the critical requirements are properly identified.

- Identified the limitation of the required hardware and software.
- Discussions with team members and supervisors to come up with an idea on available requirements.
- Define the requirement and scope of the system based on gathered requirements and feasibility study.

**2. System design**

Before starting the implementation process, we need to have a clear idea about how this application will help to the organization and the security analysis. System design can be prepared by using the data gathered in requirement specification phase. Overall architecture of the system and hardware specification are defined with the help of the system design. In this phase, we must identify the features of the four main modules in the system in order to build up the system flow, user interfaces. We can conduct a system design review to make sure that all the requirements previously gathered are addressed in the system design since it is important to defined the system before the implementation phase.

**3. Implementation**

Implementation of the application is decided to carry out as three separate process as given below.

- Database implementation.

  To keep the record of the clients (probes), store the alerts, and traffic data that send by the clients we need have well-organized database. Implementation of the database important in our implementation process since it stores core information of the system.

- User interface implementation.

  In order to provide easy, understand and visualization about the received alarms, live information on activated clients (probes), alarm types, search recorded data and client configuration our goal is to design user friendly interfaces during interface implementation.

- Coding implementation.

  Coding will be done using python programming language. Coding implementation will be great challenge since we have to extract data through web request and data transmitted to removable devices for analyzing processes. In addition, we have to customize selected machine learning algorithm for system's analyzing process.

**4. Testing**

Goal of the testing phase is to identify any defects in the system before handing it over to the user. Testers check whether the system achieved all the requirements defined in the requirement phase. Testers have to follow several levels of testing before release a product to the customer.

- **Unit testing**

In this level, Each and every sub unit is tested separately, and it helps to make sure individual units work properly without any defects**.**

- **System testing**

Performance of the whole system is tested in this level. System testing checks functional and nonfunctional requirements of the system**.**

- **Acceptance testing**

After unit testing and system testing, Product is handed over to customers to test the functionalities of the system and get the acceptance for the system.

### 3.2. System functions

#### 3.2.1. Email Analysis

Main method of communication in any corporate environment is Email, in each and every day thousands of emails are exchanged within the organization and from organization to the outside. Each one of these mails could be the email that leaks a very important data of the organization to an unauthorized party voluntarily or involuntarily. Due to the massive number of emails that are exchanged within a day it is the major vector that has the potential to leak data from an organization [13].

Current Data leakage detection systems use text-based detection systems such as keyword-based searches to identify potential data leakages from emails and they are capable of identifying text on images that are embedded in email or attached as attachments to the mail [14]. But they lack the ability in identifying a potential data leakage when the keywords are removed from the email or the attachment. A potential malicious insider could craft the email in such a way that keywords are removed or unclear so that optical character recognition modules fail to identify the words. This attack will have a great chance of success with current systems. Another area where keyword-based searches fail is when it is up against diagrams, a corporate environment could have so many diagrams that are very important like design diagrams, blueprints which are highly critical data for a organization. So, identifying any leakage of these types of diagrams is also very important.
Aleksi is performing the email analysis at the end point and it monitors all SMTP outgoing traffic. Emails that are exchanged within the company are ignored from monitoring and it concentrates on emails that are going out of the organization. This is achieved by looking at the domain part of the recipient address.

As the first step, A will check the email body for any sign of data leakage using the text analyzer module, and then it will check the body for embedded images and analyze them using the image analyzer module if any image was found.

Next it will download the attachments of the email and look for patterns of data leakage using its' document classification algorithms.
If any potential data leakage is detected during any of these steps email and its attachment will be send to central server as an alert with additional information including time and date of incident, severity assigned by tool, type of data classification rule that the alert was triggered, username and computer name of the sender.

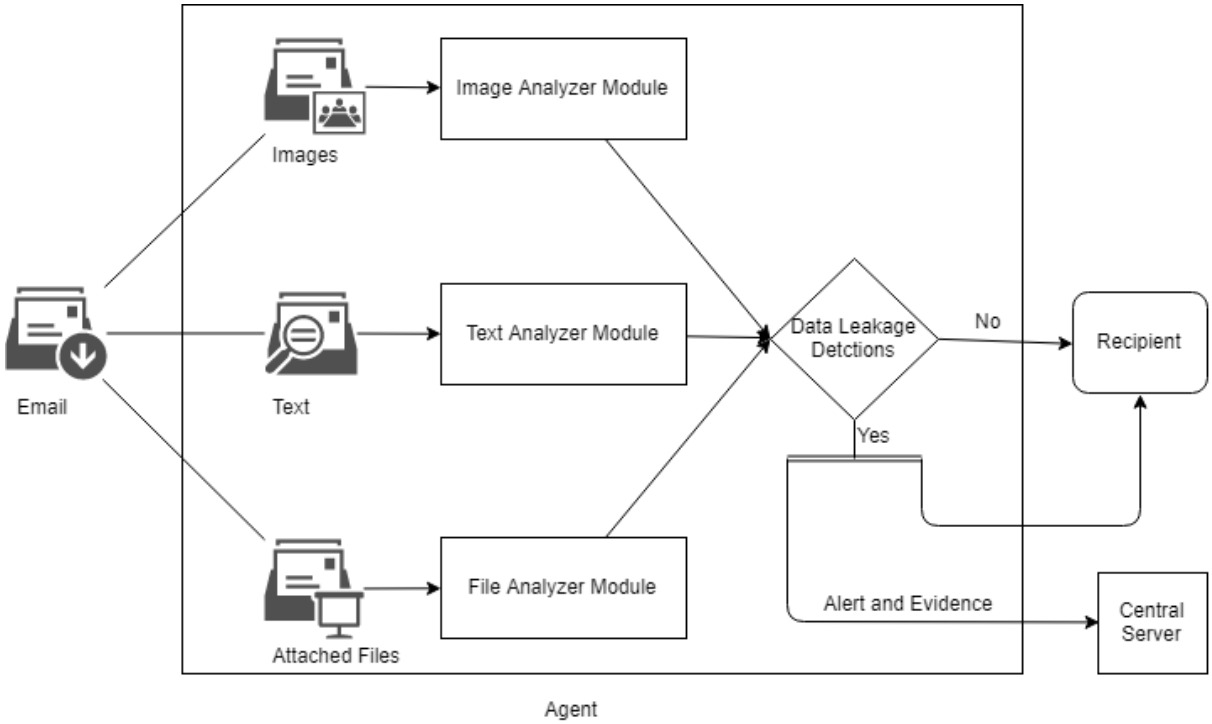


Figure 4 – Email analysis

### 3.2.2. Web Post Analysis

All most all the modern business activities rely on Internet due to the easiness and efficiency of information exchange. Even though internet provide huge number of benefits to organizations, it causes organizations to face information security risks. Insider threats among them poses the most danger to organization. One of the major threat faced by the organization is data leakage. With the internet availability, malicious insider can easily send or post organization's sensitive data on the web [15].

The solution we propose is a web post detection system design to deploy in organization's hosts to analyze, detect and send alerts to main system when sensitive data leave from hosts via internet. The primary objective of this component is to alert the main system when sensitive data leave from the organization.

Most of the existing detection systems use anomaly and text-based pattern detection mechanism to identify data transmit from devices which is produce large number of false positive alerts [15]. The goal is to combined machine learning techniques to improve the efficiency of the detection mechanism to reduce false positive alerts.

There exist two strategies, which we will used on web post detection system.
1. Develop a browser plugin to capture and analyze data transfer to internet. This approach allows monitoring all the web requests and responses process with in the browser before traffic is get encrypted. This will give opportunity collect the unencrypted web traffic, which will lead to reduce analysis workload.

2. Develop a proxy to capture and analyze data transfer to internet.
This approach allows web post detection system to capture and hold the web traffic, which provide capability to stop the traffic if sensitive data detected during analyzing process.

The traffic collected by one of the above methods then stored locally to use for analyzing process. System will design to analyze both text and images using machine-learning algorithms. If sensitive data detected during the analyzing process alerts will send to central server and then detected data will forward to the central server for later analyzing and confirmation process. Remaining traffic data recorded in locally will archived and remove periodically.

After the implementation of the suggested, web post detection system will work according to the steps given below.

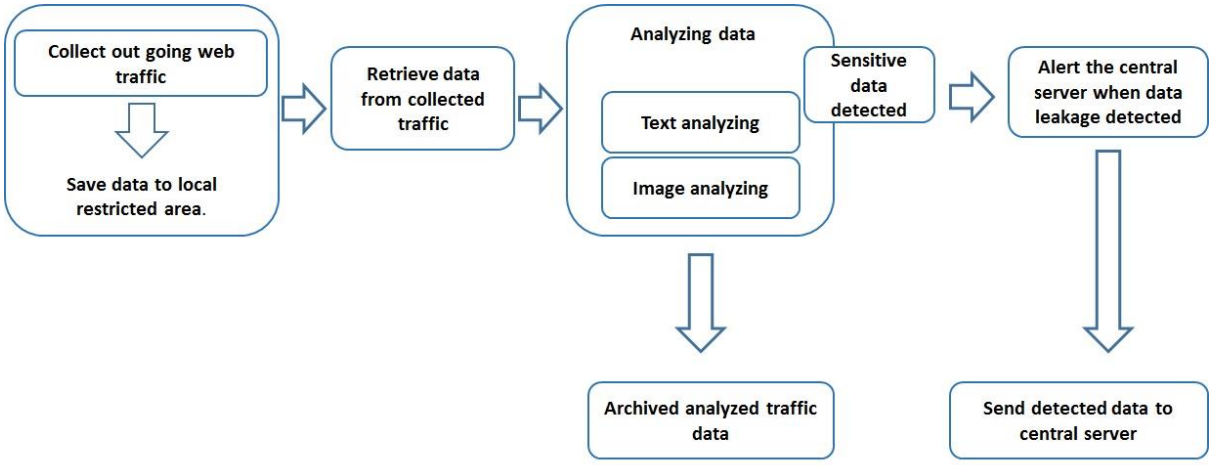


Figure 5. The methodology in which web post detection system operate.

### 3.2.3. Removable Device Analysis.

Universal serial bus (USB) storage devices are very useful medium when transferring data. However malicious insiders / unintentional acts of employees can be leads to organizational confidential data exfiltration and there's various way of doing so, this section mainly focused on preventing confidential data exfiltration using USB devices [16]. Continuously monitoring and analyzing the content transferred into the USB storages is a must. Otherwise unintentional or insider threads might lead to confidential data leakage of the organization. The main target establishes a method to effectively monitor USB traffic and generate alerts where the confidential data has leaked while still allowing employees to access USB devices and transfer media which are not confidential for the specific organization [16]

There're various windows-based tools used for USB auditing such as Windows Task Scheduler, Trend Micro's OSSEC host-based intrusion-detection system (HIDS), the Splunk log analysis engine where IT administrators can centrally monitor, audit USB devices' usage and determining whether sensitive data has been leaked or not. In our DLP solution, we're trying to automate that process using Machine Learning algorithms [17].

The first thing we have to do is capture the traffic transferring into USB storages. There Are two ways to intercept this transmission,

1. Hardware Interception
2. Software Interception [18]

Since we're focusing on Application level DLP solution we're using the second method. There Are two strategies to implement,

1. Actively monitoring USB traffics – Here we disable USB write permission [19] on the windows host and if anyone wants to send an attachment to a USB storage, it can send by using file sender application embedded into Aleksi. This way it holds the USB traffic and sends it to well-trained machine learning algorithms to analyze for confidential data exfiltration. If the result says it cleans, then Aleksi allows the user to send the file. If the result says it contains confidential data, then Aleksi drops the transaction and create an alert which contains the sender's information and the file attachment, and then Aleksi to the central server to take further action.
2. Passively monitoring USB traffics – Here Aleksi will not keep holding the traffic [20]. Aleksi will be analyzing the traffic passively and send alerts to the central server.

Depends on implementation and efficiency we'll choose one method mentioned above.

Here's how data exfiltration on USB devices' detection part going to work,

1. First, as shown in figure 1, the Aleksi will capture the transaction in the middle and decode it.
2. Then Aleksi sends to the evaluation section where evaluate for data exfiltration.
3. Well trained text-based analyzing algorithms and image classification algorithms will be analyzing the traffic and generate alarms if any policy violation happens.
4. If the file is clean, Aleksi continues with the sending process. If the file has confidential data in it, the Aleksi will ban the user from sending attachments to USB medium and send the alarm along with the file which detected as true positive.

### 3.2.4. Cloud storage upload Analysis

Cloud computing is one of the fastest growing field in IT industry due to its services. Organizations mainly use cloud, platform as a service and storage as a service [21]. Cloud provides virtual resources and it helps to reduce the hardware cost well as physical maintenance cost and technical issues. Even though cloud computing is very efficient and cost effective it creates security and privacy risks. Attacks to the cloud can occur from insiders. malicious users can upload sensitive data to private clouds.

Since transferring data to the cloud in large scale it can create a risk of data breach. So that there should be proper access control and sensitive data detection policies in an enterprise environment [22] Most organization use auto sync feature to backup files to the cloud. Another challenge in cloud uploads is that Cloud Synchronization process. Most of the cloud service providers provide features to auto synchronize a given directory with cloud. In our proposed solution we analyses application logs to detect sensitive data auto synchronizing with the cloud. [23]

In the proposed DLP solution, our approaches to detect direct sensitive data upload to the cloud through browser and synchronization is Using a proxy to capture data and analyze them to detect any sensitive data uploads. One of the advantage of using proxy over plugin is that, it can be using to detect and prevent.

## 4. Description of Personal and Facilities

| Name | Student ID | Functionality |
|---|---|---|
| R.A.C.D Ranathunga | IT14030918 | ● Removable Device analysis<br>● Text Document analysis |
| A.R.Jayaweera | IT15084064 | ● Cloud storage upload analysis<br>● sensitive .xlsx and .xls documents |
| S.P.D Kaveendra | IT15017598 | ● Web Post analysis<br>● .doc and .docx analysis |
| Weerasuriya.A.A.H.W | IT15038388 | ● Email analysis<br>● Image analysis |

Table 2 : Personal and Facilities

## 5. Budget and budget Justification

| Description | Price(Rs.) |
|---|---|
| Model training using cloud | 17000.00 |
| Printing cost | 4000.00 |
| **Total** | **21000.00** |

Table 3 : Budget

Since we are using machine learning to develop the Aleksi system, there should be a platform to train the algorithm and we are using cloud platform to train our data classification algorithms.

## 6. Conclusion and Recommendations

In conclusion Data loss prevention (DLP) is a technique used by security analysts in order to ensure that employees do not send confidential data and information outside the enterprise network. Furthermore, there are four main DLP solutions, they are,

- Network-based data loss prevention (DLP) solutions
- Datacenter or storage-based data loss prevention (DLP) solutions
- End-point based data loss prevention (DLP) solutions
- Content-aware data loss prevention (DLP) tools

Researches show the main focus of enterprises is to prevent data leakage and theft, therefore this research will be based on the 3rd DLP solution, which is the End-point based data loss prevention solution. This solution is mainly focused on end user activities, in other words this is where the DLP system will be observing PC-based systems such as laptops, desktop computers and all other end user devices used within the organizational network, for example, transferring confidential data from one device to another, leaking data via social media, email communications, printing documents and more. Such solutions are event driven where the endpoint user monitors the network for suspicious actions. Furthermore, these solutions can be programmed in two ways such as passive monitoring mode or to active monitoring mode where the agent will bar certain activities from users.

However, there are many DLP solutions in the market that claim to do all of the above, hence majority of the DLP software's have a considerable percentage of false positivity, therefore this DLP system will be programmed to apply machine learning techniques in order to decrease the chance of false positivity, increase user friendliness, efficiency and produce a highly accurate DLP system.

# References


[1] Ponemon Institute LLC, "A Websense® White Paper The ROI of Data Loss Prevention (DLP)," May 2007.

[2] B. Hauer, "Data and Information Leakage Prevention Within the Scope of Information Security," *IEEE Access,* vol. 3, p. 2554–2565, 2015.

[3] E. Costante, D. Fauri, S. Etalle, J. D. Hartog and N. Zannone, "A Hybrid Framework for Data Loss Prevention and Detection," in *2016 IEEE Security and Privacy Workshops (SPW)*, 2016.

[4] B. Hauer, "Data Leakage Prevention - A Position to State-of-the-Art Capabilities and Remaining Risk.," in *Proceedings of the 16th International Conference on Enterprise Information Systems*, 2014.

[5] G. Raho, R. Al-Shalabi, G. Kanaan and A. Nassar, "Different Classification Algorithms Based on Arabic Text Classification: Feature Selection Comparative Study," *International Journal of Advanced Computer Science and Applications,* vol. 6, no. 2, 2015.

[6] A. Patra and D. Singh, "A Survey Report on Text Classification with Different Term Weighting Methods and Comparison between Classification Algorithms," *International Journal of Computer Applications,* vol. 75, no. 7, pp. 14-18, 2013.

[7] L. Martin, "Understanding DLP," Information Security Today, [Online]. Available: http://www.infosectoday.com/Articles/DLP/Understanding_DLP.htm. [Accessed 24 March 2018].

[8] T. Torsteinbø, "Data Loss Prevention Systems and Their Weaknesses," University of Agder, 2012.

[9] Info-Tech Research Group, "Vendor Landscape: Data Loss Prevention," Info-Tech Research Group, 2016.

[10] J. Brownlee, "A Gentle Introduction to Transfer Learning for Deep Learning," 12 December 2017. [Online]. Available: https://machinelearningmastery.com/transfer-learning-for-deep-learning. [Accessed 21 March 2018].

[11] A. Quattoni, M. Collins and T. Darrell, "Transfer learning for image classification with sparse prototype representations," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference*, Anchorage, AK, USA, 2008.

[12] X. Xia, C. Xu and B. Nan, "Inception-v3 for flower classification," in *Image, Vision and Computing (ICIVC), 2017 2nd International Conference*, Chengdu, China, 2017.

[13] F. Y. Rashid, "Email Main Source of Data Leaks in Organizations: Survey," Eweek, 11 September 2011. [Online]. Available: http://www.eweek.com/security/email-main-source-of-data-leaks-in-organizations-survey. [Accessed 23 March 2018].

[14] P. Zilberman, S. Dolev, G. Katz, Y. Elovici and A. Shabtai, "Analyzing group communication for preventing data leakage via email," in *Intelligence and Security Informatics (ISI), 2011 IEEE International Conference*, Beijing, 2011.

[15] Q. Li, L. Zhu, W. Shang and S. Zeng, "CloudSync: Multi-nodes Directory Synchronization," in *2012 International Conference on Industrial Control and Electronics Engineering*, 2012.

[16] P. Walters, "The Risks of Using Portable Devices," 2012.

[17] J. Silowash and T. B. Lewellen, "Insider Threat Control: Using Universal Serial Bus (USB) Device Auditing to Detect Possible Data Exfiltration by Malicious Insiders," January 2013.

[18] D. Elmi, "Sniffing USB Traffic - Different Approaches," Dlog, 22 October 2012. [Online]. Available: http://dan3lmi.blogspot.com/2012/10/sniffing-usb-traffic-different.html. [Accessed 29 March 2018].

[19] M. Huculak, "How to enable write protection for USB devices on Windows 10," Windows Central, 10 November 2016. [Online]. Available: https://www.windowscentral.com/how-enable-write-protection-usb-devices-windows-10. [Accessed 23 March 2018].

[20] WireShark, "USB capture setup," WireShark, 10 March 2017. [Online]. Available: http://wiki.wireshark.org/CaptureSetup/USB. [Accessed 27 March 2018].

[21] N. P. Doe and S. V., "Secure service to prevent data breaches in cloud," in *International Conference on Computer Communication and Informatics*, 2014.

[22] Y. J. Ong, M. Qiao, R. Routray and R. Raphael, "Context-Aware Data Loss Prevention for Cloud Storage Services," in *2017 IEEE 10th International Conference on Cloud Computing (CLOUD)*, 2017.

[23] Lee, F. S. Alamri and K. D., "Secure sharing of health data over cloud," in *2015 5th National Symposium on Information Technology: Towards New Smart World (NSITNSW)*, 2015.

[24] "Machine Learning Sets New Standard for Data Loss Prevention," [Online]. Available: http://eval.symantec.com/mktginfo/enterprise/white_papers/b-dlp_machine_learning.WP_en-us.pdf. [Accessed 14 March 2018].

[25] A. Vaidya, P. Lahange, K. More, S. Kachroo and N. Pandey, "DATA LEAKAGE DETECTION," *International Journal of Advances in Engineering & Technology,* 2012.

[26] K. W. Kongsg˚ard(B), N. A. Nordbotten, F. Mancini and P. E, "Data Loss Prevention Based on Text Classification in Controlled Environments," Norwegian Defence Research Establishment (FFI).

[27] L. Shi, S. Butakov, D. Lindskog and R. Ruhl, "Applicability of probabilistic data structures for filtering tasks in Data Loss Prevention Systems," in *29th International Conference on Advanced Information Networking and Applications Workshops*, 2015.

[28] Mellouk, B. Augustin and Abdelhamid, "Traffic Patterns of HTTP Applications," *2011 IEEE,* no. Dec 2011, pp. 1-6, April 2011.

[29] M. Hart, P. Manadhata and R. Johnson, *Text Classification for Data Loss Prevention,* Computer Science Department, Stony Brook University.

[30] R. Tahboub and Y. Saleh, "Data Leakage/Loss Prevention Systems (DLP)," in *2014 World Congress on Computer Applications and Information Systems (WCCAIS)*, 2014.

[31] Y. Shapira, B. Shapira and A. Shabtai, "Content-based data leakage detection using extended," in *Dept. of Information Systems Engineering, Ben-Gurion University of the Negev*.

[32] CISCO, "Data Leakage Worldwide: The High Cost of Insider Threats," in *CISCO white Paper*.

# Appendix A: Gannet Chart for ALEKSI System