Data Leakage Prevention System: A Systematic Report

Sheela Gowr. P, Kumar. N

Abstract: ***Technology is growing exponentially and in the fast growing technology more fast processing devices are being introduced. A mobile phone could even handle high processing software and even a small SD card could store the complete data of an organization. In this fast growing it environment maintaining security of data is highly important as a small loss of data might create critical impact in the organization, hence preventing their sensitive data became a greatest challenge. Traditionally organisation implemented methods like framing policies in an organization, implementing Firewall, virtual private network at the endpoints, but still these method started lagging as the technology developed the methodology of data leakage and data theft also attained heights. Hence there was a need for some system which could prevent leakage of data and this could be done with the help of Data Leakage Prevention system (DLPs). The DLPs are capable of detect leakage in Data at any states namely Data in rest, Data in transit and Data in use which increases the need of the DLPs.This paper briefly explains about the carious DLPs available, their limitations, providing the information about the gaps in security providing enough awareness on new developers as well as researchers and professionals for developing a next generation DLP which is capable of preventing data loss.***

Keywords: Data Loss, Data Protection, Data Security, Data Leakage detection, Data Loss Prevention.

I. INTRODUCTION

Most of the companies have moved to cloud computing as it gives more advantages such as remotely working, easy of sharing files and making day-to-day activities simpler, the biggest challenge for the organization is data leakage. Even though cloud computing has many advantages it comes with higher risks for those organisations[1]. The most critical threat of the organisations which moved to the cloud computing is loss of their personal and sensitive information deliberately or inadvertently. Most of the organisations have more clod data breeches as they does not leverage best practices. Data breaches are not occurring after the swift of data to digital model, it was happening even when the data are being saved manually as papers in files[2]. When the data are not digitally available an unauthorized person looking in to the files without proper permission or looking into private information which are not properly destroyed are also called as Data breaches. As data breaches frequency increased rapidly through 1980s, 1990s and early 2000s organisations and public awareness on the potential harm caused due to it also rapidly got increased[3].

Revised Manuscript Received on November 15, 2019

Sheela Gowr. P, Research Scholar Department of Computer science and Engineering
Vels Institute of Science Technology and Advanced Studies
Kumar. N, Associate Professor, Department of Computer science and Engineering Vels Institute of Science Technology and Advanced Studies
sheela.se@velsuniv.ac.in, kumar.se@velsuniv.ac.in

The four main streams of Data Breaches are Ransomware, Malware, Phishing and Denial-of-Service (DoS). Even data breaches are occurring prior to 2005, many biggest data breaches reported in history in 2005 or beyond it[4]. As every data are being recorded in cloud the volume of Data across the world are keep on increasing day after day which gives cyber criminals bigger opportunity to expose a big volume of data in a single data breaches attempt. The Privacy Rights Clearinghouse [5]has reported that in 2005 alone 136 data breaches have occurred and 4500 data breaches have also made publicly occurred since 2005 and 816 million records belonging to individuals have also been breached.

Many expert committee members, analytical agencies and media have been continuously attempting to bring out the biggest Data Breaches out in the past, since 2005 the Data breaches happening in United States are being continually being reported by Statistic's reports states that Data breaches are being in on upwards side since 2005 every year[6]. 157 breaches have been reported in 2005 with 66.9 million records are being exposed, whereas in 2014 it was 783 breaches with 85.61 million records being exposed which was 500 percent more when compared with 2005. In 2017 the breaches reported was increased to 1,579. All these reports are only from United States and they are reported officially and this may increase while considering which was unreported. While comparing with Verizon DBIR or other leading industry standard breach reports Statista's reports will be more conservative.

The number of breaches reported in United States is being increasing every year and it was reported to be 614, 783, 1093 and 1579 breaches in the year 2013, 2014, 2016 and 2017 year respectively. Accordingly to the reports of 2018, the cost per record last in a breach was found to be 148 US Dollars[7]. The average cost of per breach was found to be 3.86 million US Dollars. Even a breach in a small business with just a loss of 1000 records would cost few thousand US Dollars.

Companies, organisationsand agencies not only have primary information in cloud but may possess high sensitive information such as medical records of patients, trade secrets, and military classified documents. It will be more critical if these sensitive data are leaked to unauthorized persons or agencies[8]. Some of the basic strategies to avoid data leakages are encrypting such sensitive data, change the password of the cloud server very often, train the staffs on safe handling of Data, and also set permissions for the employees[9]. Every organisation started looking for a solution to reduce the Data breach occurrences and as a result Cloud based Data Leak Protection methods are rapidly increasing day by day[10].

Gartner estimation reports that by the end of 2019 approximately 90 % of the organisation will own a DLP solution to protect Data Breaches which was actually 50% higher than 2016. Many vital techniques has been created to build a strong DLP[11], which keeps the server safer and breach protected.

Cloud based DLPs could be broadly classified into two categories and they are Integrated DLP and Enterprise DLP. Enterprise DLP monitors the email and the network traffic of the firm and tries to reduce the data leakage[12]. On the other side integrated DLP provides secured gateway policies, encrypted solution for email, secured gateways for email which scans the incoming and outgoing emails of the firm, protected ECM environments, better data classification solutions and utility discovering.

Providing solutions for the serious issues, security experts put enormous efforts to develop an advanced DLP. In the past three decades solutions namely, intrusion detection systems (IDSs), intrusion prevention systems (IPSs), and virtual private networks (VPNs) have been introduced which was proven to be providing satisfying results when the data are well defined, structured and constant[13]. This measure was able to provide a simple solution to small extent. For example script rules written in firewall could definitely block access by unauthorized persons but the same data parts could be accessed by other means such as instant messaging (IM) or email attachments. Thus security measures including firewall, IDs, and VPNs lack in security and advanced attacks made on data servers[14]. Hence as a result to overcome the deficiency advanced DLPs have to be designed which must be capable of protecting confidential data and prevent data being misused by applying predefined rules. The DLP must be a better solution than the existing security solutions. Researchers and industries are also putting extreme efforts to provide an improved solution as the past solutions are still lagging.

II. DATA LEAKAGE PROTECTION (DLP)

The reports of Forrester Wave[15], most of the DLPs designed earlier was focusing on network points through which the sensitive data are being left out from the organization. In the next stage started monitoring removable storage devices such as USB drives or External HDDs, through which the data leakage could be happen[16]. Later DLP solution could start focus on detecting data leakage by copying the data to secondary storage devices at the endpoint even when they are not connected to the network.

The most important task of the DLP is protecting the sensitive data which are available in unstructured form (e.g., Source code of a project, customer information's, design of a product) hence DLP solution designers and providers started to design a better solution using approaches based on fingerprinting and natural-language processing[17].

Most important feature of the DLP is providing a centralized admin system which holds the entire control of the sensitive data and Specific rules and procedures when an attempt for data leak is detected[18]. The rules and procedures are not static and it customizable so that the users can specify what data has to be checked (e.g., email messages, web forms or data stored on sharing networks) define cardinality (e.g., more CCNs in the email) and proximity (e.g., name and mail id of a person but the other information in the email are irrelevant to the name). Some of the rules are:

1. If an Email is found to have sensitive information, the fingerprint has to be compared and it has to be redirected to the encryption gateway.
2. If a file contains the keyword "project" it should not be accessed and hence a MS-Rights Management System is to be applied to prevent accessing to the file.
3. A file containing SSNs in publicly accessed locations should be relocated to the secure server T.

The centralized admin system also adjusts itself to update the policy via DLP system modules, it also looks in to the reports on policy violation and gives alerts if necessary, it also investigates and takes action against leakage incidents[14]. Hence an important feature of the admin system will be capability of providing a detailed report of data leakage and user actions (such as keystrokes, files opened, and Web sites visited) that occurred after a leakage.

A detailed literature analysis was performed and based on it the literature can be huddled in to the following groups:

a) Detecting Misuse in information retrieval system. It deeply look in to the authorized users having permission to view the information retrieval system and are they accessing the files which they are permitted to see or accessing unauthorized files.
b) Detecting Misuse in databases. Looks in to the unauthorized access to the files by authorized users as it is believed and accepted that many data leakage are done by people with authorized access.
c) Detecting Email leakages. Data mining technology can be used which helps in automatic detection of both signature-based misuse detection and anomaly detection-based misuse. This could be done by inspecting email contents or conducting behaviour based analyses and find the group of user accounts who communicate with each other very frequently. These approaches also can be clubbed and interoperated to provide more strict enforcement.
d) Providing Network/web based protection. The usage of network bandwidth and internet keeps on increasing day by day as it could not be restricted and a result of it enterprises and organisations face a problem of keeping their sensitive data getting leaked out by the networking system. Hence a data loss prevention system is generated which handles high volume network traffic and also verify the outgoing information via the network. These systems are capable of preventing direct data leakage but not capable of identifying encrypted data leaks.
e) Providing access control. Sensitive data requires a safety environment where the data are accessed, saved and transferred. It mainly focuses weather the sensitive data are accessed by authorized user and systems.

f) Preventing data hidden in a file. Hiding sensitive data or information in a single file is not safest method as using the common knowledge the data can either be accessed or can be easily predicted. Hence sensitive information has to be saved in multiple documents and multiple locations in a network.
g) Detecting malicious insiders using honeypots and honey tokens. In this honey tokens are kept inside the applications of the organisation. Later these take permissions from users and monitor the user actions and find the true intension and reveal it.

DATA LEAKAGE PREVENTION SYSTEMS

Enterprise content aware DLP solution or DLP solution in general offers many approaches to monitor, protect sensitive data at their end points[19]. Generally these types of systems validate and provide authorization to applications whenever access or transfer of confidential data is required. These systems also takes care of monitoring clients endpoints and the network traffic in the clients environment, they also prevent copy and paste operation on sensitive data and taking screenshots can also be prevented. Present DLP are able to have full control over the access of confidential data and access of it by the user[20]. Apart from many issues and limitation it tries to prevent unauthorised access by any users or applications.

Many standard functionalities are found common among the DLP Solutions namely preventing data transfer to external storage devices, preventing screenshots, preventing printing files and also scan encrypted data and find the hidden files[18]. The market leaders provide different DLP systems each having different behaviours and functionalities and few of the market leaders their DLP function, advantages and limitations of them are briefed out in Table 1 and a typical DLP is shown in Figure.1.

Table.1 DLP Functionalities, advantages and limitations of market leading DLP Providers

DLP Provider	DLP Functionalities, advantages and limitations
EMC – RSA	• EMC offers the "RSA Data Loss Prevention Suite" (RSA DLP) • RSA DLP perform scan on client endpoints, share point products using the keywords and regular expressions • Its capabilities are limited to IP version 4 and 6 and higher level protocols namely HTTPS and SMTPS • This suit does not support mobile device management (MDM), Linux based OS and Cloud infrastructures.
McAfee	• "McAfee DLP Monitor" makes use of a switched port analyser (SPAN) port or a network tap which keep on monitor the network traffic, the data accessed, sender and the destination • "McAfee DLP Prevent" blocks or redirects the network traffic but limited to email traffic per appliance namely 30 concurrent connections via SMTP and 4000 ICAP connections. • "McAfee DLP Discover" classifies well organised data, based on the metadata values and filename extensions • These solutions are specialised in Microsoft applications and does not provide support to Linux based OS. • Support for Mobile Device Management (MDM) are offered separately for different mobile devices
Symantec	• "Symantec Data Loss Prevention 12" can be installed on both operating systems of Microsoft and Linux as well as on Microsoft Windows Server operating systems • Symantec DLP consists of three products namely "Symantec Data Loss Prevention Network Monitor", "Symantec Data Loss Prevention Network Prevent for Email" and "Symantec Data Loss Prevention Network Prevent for Web" • Its support is limited to specific service providers of cloud systems and social media • Unlike using keywords and regular expressions it deploys vector machine learning techniques for building statistical models which are based on training the system with positive and negative example documents • These DLPs also tracks the file usage which helps in enforcing access rules and understanding leakage incidents.

Verdasys	• "Digital Guardian (DG) version 6" specializes in unstructured data and extended operating system support. • Verdasys offers different network agents, such as "DG XPS DIRECT", "DG XPS MAIL", "DG XPS WEB", and "DG NETCOM", • With the help of agents they takes control over context-based data monitoring and classification. It also enforces data policies on Windows, Linux, Mac OS, VMware, Citrix, Hyper-V, BlackBerry Enterprise Server, Exchange ActiveSync, and iOS platforms. • Unstructured data are identified and classified according to context parameters by considering user, application, and activity such as the creation, access, revision, or transmission.
Websense	• Websense combines the "Websense Data Security Suite", "Websense Email Security Gateway Anywhere", and "Websense Web Security Gateway Anywhere" to the "Websense Triton Enterprise Suite" • It uses different agents and scans and monitors the endpoint systems of the clients. • The DLP is able to detection encryption and also looks into the geographical location of the source and the destination of the files being transmitted. • The classifiers uses support vector machine (SVM) models, and optical character recognition (OCR) methods • These capability are restricted to supported network traffic such as e-mail, web, and IM protocols.

Advantages of Cloud Computing Data Leakage Protection Solution

The major advantages cloud computing data leakage protection system includes: scan and audit the information on the cloud, Discover confidential Data in the cloud, automatically applies controls to prompt, block, and encrypt the sensitive data and Hold the data visibility and controls required to comply with security standards[21].

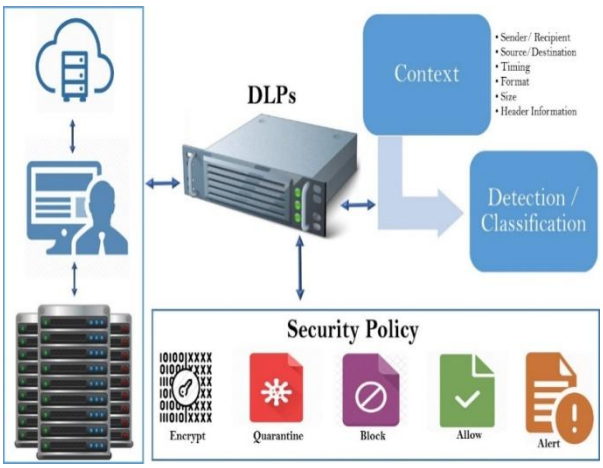

Figure.1 Typical Data Leakage Prevention System

Data States

Data will be either in one of the three states namely Data in use, Data at rest, and Data in motion and all the functionalities will be characterized based on the state of the Data. Protection of Data in rest could be done by content discovery Solutions[9]. These products scan all the location where Data are saved which includes laptops, data servers, email servers, entire database system and document management systems in order to detect sensitive data which might be stored in unauthorized locations. The two basic techniques for content discovery are:

Local Scanning: In this method a local agent will be installed in the machine and scans all the files on the host which violates the policies, if such issues are found actions including relocation, encryption and quarantine are performed[11]. The memory and processing power of the target systems are becoming limitations of the system as it is totally depended on the host systems, being installed locally its main advantage of the system will be the system is always active even if they are disconnected from the network.

Remote Scanning: Using File sharing applications a connection to the server is made and scanning process is carried over[18]. The entire process is carried on over network, hence based on network traffics and limitation in bandwidth makes the scanning process slower which will be a major limitation of the system.

Protection of Data in rest can also done by encrypting the data which are available in the end point. This could be accomplished by full-disk encryption process with the help of access control[22]. This could protect sensitive data. For example if a laptop containing sensitive data is lost or stolen the sensitive data cannot be stolen or misused.

Protection of Data in use are done by local agents that locally monitors and prevents data leakage including copy and paste, printing of data, taking screen shots, copying of data to external storage devices such as USB/CD/DVD/External HDD, unauthorized data transmissions and use of data by unauthorized applications.

Protection for data-in-motion is established by means of a network-based solution which searches for contents violating the policy[23]. These products are deployed in the endpoint of the client systems which captures the entire packet and conducts full scan on it. These products inspect various protocols (HTTP, FTP and IM), Email attachments and suspicious emails are quarantined or filtered.

Deployment

On the basis of the type of the targeted data, DLPS systems are deployed in many forms. The Data in use requires a local agent to be installed in the endpoint of the system which prevents transfer of data through portable media. It also restricts users from taking print copies, screenshots, and audits all reports related to the confidential data access[24].

For the Data in transit DLP with special processing capabilities has to be used. They take care of the network traffics, acts like proxy servers, and alerts administrators about unauthorized access or usage of sensitive data. It is also capable of coordinating with other major mechanisms like SSL proxies and endpoint firewalls[25].

III. CHALLENGES IN DATA LEAKAGE PREVENTION SYSTEMS

For "Data in rest" and "Data in use" confidential data are leaked through secondary storage devices which could be resisted by implementing host DLPs but the other network channels such as email or IM are still available through which the data leakage are still possible[13].Logical access rights could be imposed to avoid users accessing sensitive data, but still the same data could be accessed by taking print out of it. "Data in transmit" could be passed through channels and extensive traffic filtering is required at the end point and various data leakage channels are shown in Figure.2.

These kind of Data leakage are the biggest threat for data security regardless the size of the organization it makes the issues more serious and causes fall in revenue to the organization. Hence every organization is likely to have their own DLP for preventing loss of data and data being leaked. The major types of data leaks include Accidental Breach, The Disgruntled or Ill-Intentioned Employee and Electronic Communications with Malicious Intent[22].

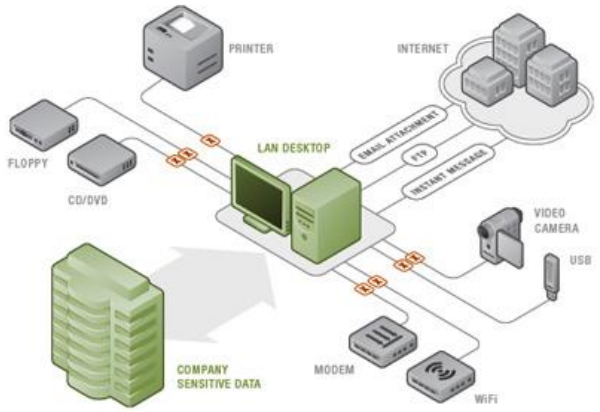

Figure.2. Data Leakage Channels

Protection Issues in Cloud Computing

The National Institute of Standards and Technology defines that Cloud Computing have three service models and Four deployment models. Cloud Software as a Service (SaaS)[26], Cloud Platform as a Service (PaaS) [27]and Cloud Infrastructure as a Service (IaaS)[10] are the service models and Private cloud, Community cloud, Public cloud and Hybrid cloud are the Deployment models respectively.

Cloud Service providers claim their solutions as more secured and reliable. But the reality is its not safer switching our products to cloud as they promise. Since 2009 almost all the cloud providers have faced several leakage accidents[28].

Security measures that have to be taken in cloud are similar to that of traditional IT environment. Cloud computing faces several risks and challenges as similar to traditional IT environment and it includes:

a. Due to its advanced features including dynamic scalability, service abstraction, and location transparency the applications on the cloud does not have a fixed infrastructure and security boundary
b. The cloud resources are owned by multiple owners and due to the conflict of interest it makes difficult in deploying unified security measures
c. Due to the functionality of shared virtualized resources, user data might be accessed by other unauthorized users.
d. Providing massive storage and fast access cloud security services also must be fast enough to handle massive information scanning in a short time

According to Gartner [4], users must enquire the vendors about seven safety measures including: Privileged user access, regulatory compliance, data location, data segregation, recovery, investigative support and long-term viability.

Data Life Cycle in Cloud

Data in cloud have to pass through the six stages and it is clearly shown in Figure.3. Tracking the data accurately helps in determining the security controls to be used in protection of data. In the first phase of the data, data are created it might be structured or unstructured[27]. It can be a word document, PDF, email information inside a database, or images. The security in this phase is achieved with the help of enterprise data security policy.

After creation of data the created data has to be stored in a secured place, hence necessary data security controls have to be implemented to have a secured environment for storing sensitive data. After this the data has to be used and hence the data has to be monitored such that weather the user or application using the data are authorised to use the data[29].

Data will be constantly shared among users and hence the data has to be monitored such that only authorized users share the data and they are properly encrypted and the sender and the recipient information are monitored and saved and the reports are forwarded to the administrator when needed[30]. At some point of time the data will not be used and it will be archived and such archived data must be scanned periodically to check any access of the data was found and if any has to be reported instantly to the administrator, finally when the volume of archived data grow beyond an extent it has to be securely destroyed as if they are leaked to any unknown persons may lead to any part of damage to the organization.

Figure.3 Data Life Cycle

Preventing Data Leakage

Data leakage could be managed with the help of various tools namely Data Leak Prevention DLPs or Content Monitoring and Filtering Tools. This is accomplished by identifying sensitive data, monitoring it and blocking the movement of such data[31]. The following measures could also be done:

- Performing deep packet inspection on the networ traffic and implementing it on networks and endpoints
- Tracking the complete session instead of analysing singular packets
- By adopting both statistical and linguistic analysis techniques
- Effectively use of specific rules and policies which prevents printing and forwarding sensitive information or data
- Implementing a single product which monitors network traffic, emails and multiple channels.
- Completely monitoring the end users activities and reporting if any unusual behaviour is found
- By encrypting all authorized communications

IV. TYPES OF DLP SOLUTIONS

A. Encryption

Encryption is a simple and effective solution which is mostly used to avoid data leakage. When a data is said to be encrypted it cannot be viewed or accessed by unauthorized persons until the data are decrypted[32]. The data is safe even if a data is encrypted and moved to external media devices and taken away from the organization environment. These measure are being necessary as usage of external storage devices are becoming unavoidable at the fast growing IT environment. These can be applied to even smartphones and all type of removable devices.The encrypted data can be decrypted only in the system containing the encryption software which was used earlier for encryption also the user must have the encryption key[19]. The data could not be viewed in any other system which does not possess the appropriate algorithm which was used for encryption.

For network based DLPs encryption will be very challenging to be implemented. Network based DLPs tries to identify confidential data in the network and compares the data with the available data in the servers. Using appropriate encryption algorithms it makes unauthorized users very difficult to view or access the sensitive information or data[33]. When a unauthorized person encrypt the sensitive data and send as a mail attachment the network DLPs could not identify the data leakage which is an limitation of this type of systems.

B. Digital rights management technology (DRM)

DRM could be used where wider security is required. This is yet another system based on encryption algorithms. This system first verifies the user credentials and confirm it, later it verifies is the user having authorization for viewing or using the data and if got verified it provides access to those encrypted files[9]. The users then decrypt the files in the system and view or use the data. This method could be applied in a closed environment i.e., it can be used only when the data is within the organizations environment or within the limits of the firewall.

C. Content Management Systems (CMS)

CMS are used by the organisation for organizing digital content. CMS acts as the central control for all the content related tasks namely content creation, content management, and distribution. Hence integrating CMS and DRM gives a better solution for protecting Data in the organisation and provide assurance that DRM-CMS will provide regulatory rules, responsibility, privacy and security legislation[34].

Content Management System (CMS) has two major components and they are Content Delivery Application (CDA) and Content Management Application (CMA). CMA is a graphical based GUI application allowing users to take control over creation, modification and deletion of contents[25].CDA provides backend solutions which successfully deliver the contents that are created by CMA.

D. Scalability and integration

Like other security mechanisms quantity of data processed makes impact on the DLPs. DLPS may be of any form such as host based, network based, and storage based but it must be capable of processing the data without making delay in the workflow[12]. Factors including data size, computation capabilities and analysis technique should also be thoroughly investigated to run a best DLP system. Some of the mandatory functionalities of the DLPs are already existing in some of the existing devices in the organization such as firewalls, IDSs and proxy servers which makes DLPs to have poor association within a network[22]. A DLP could be implemented in a network only if it could process more data in a shorter time. To advance the traffic inspection process, services such as SSL proxies and VPNs should be fully integrated with DLPSs. This ensures that DLPS scan analyse traffic in plaintext.

E. Data Classification

DLPs are mostly depended on the appropriate classified data. In case of un-classified data DLPs may not efficiently perform and will it be able to distinguish between confidential and normal traffic [35]. Data classification is a commonly used procedure in military and government based applications. Military applications uses terms such as ‘restricted’, ‘confidential’, ‘secret’ and ‘top secret’ which helps in easier identification of sensitive data. Another important task of classification is to rank the sensitive level. Hence it becomes the responsible to owner of the data to decide whether it is sensitive or not, if sensitive have to rank it accordingly. But in real time most of the data owners do not take responsibilities to do those tasks which really create many uncertainties and cause deterioration to DLP tasks. If the data are not properly classified they can be leaked in the availability of DLP.

V. CURRENT METHODS

Data leakage can be avoided by implementing management procedures, market leaders are having many DLPs to suite the market need and they are already described in Table.1, but most of the existing methods have some limitations and they are also mentioned, continues Academic researchers were been held to find the right DLP and those methods are detailed in this section.

Academic Research

The term DLPs is not widely used among academic researchers. Although many DLPs have been already proposed but it has some limitations in some end and hence more research has to carried out and a complete package has to developed which suits the need of the organization in prevention of the sensitive data[36]. DLPs are classified in two streams namely, Preventive and Detective. Since the DLPs are categorised based on their types some of the existing or the trending one may be excluded.

On the preventive methods techniques namely, Policy and Access Rights, Virtualization and Isolation, Cryptographic Approaches and Quantifying are explained and in other side Data Identification, Social and Behavioural, Data mining / text clustering, and Quantifying and Limiting methods are explained in detective side and finally AI based methods have been explained with its advantages over other methods[37].

A. Policy and Access Rights:

Most of the data security software are intended to prevent the data from malicious attacks, but on the other hand the data are still under serious problems since the employees could access and share sensitive data[25]. The servers of an organisation are distributed as the number of employees working from remote keep on increasing, hence a data loss prevention policy is required to prevent unauthorized access or sharing outside the organization premises[38]. The best practices that could be implemented which could establish data protection are:

- The data for which the policy has to be implemented is identified, for this the data has to be carefully classified based on the vulnerability and risk factors.
- The people who are going to be responsible for data loss prevention are identified and they are given appropriate roles and authorization for accessing the data
- The Policy are defined starting from low vulnerable data to ward data with high risk.
- Every unit/department head of the organization should be involved in building up the policy of the organization
- All the employees of the organization must be given proper information about the organization data loss prevention policies
- Data loss prevention policies are carefully documented.
- Return on Investment of policy are determined by the data loss prevention metrics which also efficiently determine it
- The policies that are being implemented must not be a burden to the employees at any means so extreme care has to be taken while defining. (e.g., if the policy block users from attaching larger file in their email it may affects users even to share necessary data and the user have to look on an alternate for it)
- Any data leak must be reported first before blocking. So set up a data loss prevention tool to report about the sensitive data loss first and block it.

An efficient data leak prevention policy must necessarily contain the three elements namely, Location (where it is enforced), Condition (policies for data Loss) and Action (steps if a data leak is found). Hence to conclude it is observed that DLPs based on security policies are simplest measure for data leakage.

B. Virtualisation and isolation

The introduction and growth of virtualization can be effectively used in a DLP solution. This is implemented by creating a virtualisation environment while accessing the sensitive data, which helps in isolating the user activities and only trusted peoples can use those sensitive data[30].

Some of the method that could be used are: 1. creating trusted virtual domains and securely linked with each other by which the environment is off loaded and could be used which maintains the security of the data. 2. Having two different virtual machines in which one of the VM have entire access over the network and the other one is used to access data in the server and no connection will be executed between them which protect sharing of data from the server. 3. Using Server virtualization in which the resources of server is masked and the physical server is broken into many small VMs which provides a barrier for hackers to find the VM where the data is located and also the isolates the applications which are not stable since it is a threat to sensitive data. 4. Using Network virtualization by which logical virtual networks are created and help in sub-dividing network into sub-networks. This sub networks helps to reduce network traffic hence increasing the performance of DLPs[39].

In addition to the said methods some other virtualization can also be implemented namely: Session-based application virtualization, Session-based desktop virtualization, hosted virtual desktops (HVD), desktop streaming, Local desktop virtualization, sandboxing, and virtualswitching and server isolation.

C. Cryptographic Approaches

Cryptography is a normal approach of protecting data from unauthorised access. Cryptographic methods have been proven as high end safety measure for unauthorized access. In this method the original data are encrypted using a key and hence without the proper key it could not be decrypted back[32]. Thus encrypted data could be transferred via common networks which are also safe since if any unauthorised user takes the data it cannot be accessed without the right encryption key.

In Cloud infrastructure the user might not have actual or physical control over the information stored in it, so one of the best methods that could protect the data in cloud will be data is stored cryptographically with user having the cryptography key with them. Also experts suggest that information in motion and information at rest are best protected by cryptographic security measures[40].

Most of the organizations encrypt the data before moving them to cloud this method is more secured as the data is completely encrypted before they leave the organization environment and the data can be only accessed with the help of decryption keys. Normally cryptography may be applied in few methods: if a user is asked for a key for opening the encrypted data the user have to provide the key to access the data in case of wrong entries the user could not view the data and also it will be informed to the admin[25]. In other case the user may be provided temporary single time key for accessing the information and after one time usage it will be disabled. Among both of the methods the first one is suitable for stand-alone machines and the second one is suitable for remote access.

But this method is more secure only if the organization possesses an effective file system.

D. Quantifying and limiting

Activities namely surfing the web, running a specific process, accessing a particular file can also lead leakage of sensitive data which grant access to unauthorised persons. In some cases the data are not directly released, it happens when programming is done using very simple programming language[41]. In that case quantity of sensitive data is identified. Hence a quantifying approach will be helpful in identifying the quantity of sensitive data and also tries to check the leakage rate in bits. Although these method seems to be good in analysing the data it has many limitations such as: they are unable to filter out data of uniform resource locators, also could not find data available which is not written in simple imperative language. These methods are very closer to secure publishing which provides tools to publish information and on the other side maintain the privacy of the information[28].

E. Data identification

Like most of the Antiviruses and Spam filtering most of the DLPs do deep packet inspection to find weather data is getting leaked in the network. But to execute these type of methods they require previous documents including finger prints, regular expressions and partial or full data[15]. These systems scan through the networks and scan every data and compare the data with the existing one and if a match is found it is immediately blocked and reported. This system has its one limitations too, since deep scans are taken place it take more time to handle the network forming traffic in the network and could not handle network with high traffic rates also similarly the entire system focuses data matching and hence if the data is modified and made to leak the system could not detect the leak as its partially or completely modified. With these limitations the system produces low false positives.

F. Social and Behaviour analysis

This method is focused on making enough analysis in the interaction between different users of the organization. Every user in the organization are considered as nodes and all the interactions and communication between the nodes are completely monitored and relation between the nodes are drawn. It also scans the type of conversation and thesize of the transaction and tries to predict the relation between the nodes, and if any misbehaver found will be reported immediately[42]. As aware mostly human intention could not be easily predicted and hence the DLPs frequently needs its admin to look over it and see any unauthorised access in the server has been found or not. A frequency line is also drawn between nodes having similar communications or common topic discussions hence identifying the nodes into groups so that it would be easier to identify the nodes having different behaviour. The main intend of these type of DLPs is to detect the people have different intend than others who tries to take access over specific data or transfer those data to others making a leak in the network. One of the other common method is creating honeypots in the server which makes hackers and unauthorised persons fall in the trap created by us once they start accessing the honeypot the activity will be reported and hence they could be easily identified and blocked[43].

G. Data mining and text clustering

Most of the DLP discussed will work for the case where data are properly classified and for non-classified data we need a different DLP that can scan through network and find out the data leakage. For these situations data mining techniques works good as it scans through the big documents and identifies keywords, recognize complex patterns and take decisions based on this hence these systems could be related to machine learning systems[44].
These type of DLPs mostly will have two phases namely training and execution where in the training phase sample documents are provided in which the DLPs conducts the scan and recognize complex patterns and also make classification patterns. Once this is ready the DLPS starts functioning and if a data is found it is scanned and matched with the possible cluster and identify the containing class from which it will be easy for detecting weather the data is sensitive or not[4].
Text clustering method based DLPs are also developed. They are meant to perform well in natural language processing systems but could not perform well in the DLPs as they involve artificial intelligence based detection process and hence required more updates to be carried over. These methods can also be efficiently used to find out the leakage happening in social networking sites. Being an developing phase it has increased false positives[19].

Key focus area for DLP

As the technology is keep on updating the data theft ways also gets updated and hence every organization must keep on updating their DLPs to get away from data leaks. Table 2 shows a clear picture about the various area that has to be focused for an effective DLP for various data states.

Table. 2 Data states and their Focus Areas for effective DLPs

Data State	Focus area
Data in Motion	• Perimeter Security • Networking Monitoring • Internet access control • Data collection and exchange with third parties • Instant messaging • Remote access
Data in Use	• Privileged user monitoring • Access/usage monitoring • Data sanitation • Use of test data • Data redaction • Export/save control
Data at Rest	• Endpoint security • Host encryption • Mobile device protection • Network/intranet storage • Physical media control • Disposal and destruction

Advantages and Limitations

Every method mentioned above has their own advantages and limitations and they are described clearly in Table.3. Further research has to be carried over to rule out the challenges and thus developing a DLP system that prevents Data Leakage by any means

Table.3. Strength and Limitations of DLP Methods

Method	Strength	Limitations
Policy and Access Rights	✓ Suitable for well-established organization ✓ Easy implementation process ✓ Best for Data in use and at rest	× Could not be implemented for unclassified data × Could be affected by present policy standards × It is prevention method and detection could not be done
Virtualisation and isolation	✓ Mandatory Hardware implementation ✓ Does not require regular admin interference ✓ Existing classification could be used	× Not matured when comparing with other methods × It makes more hurdles × It is also a prevention method and detection could not be done
Cryptographic Approaches	✓ Stronger the cryptography stronger the security ✓ Wider method with more functionalities and options	× It can secure the sensitive data but could not stop if any data leaks which is already existing × Using weaker credentials confidential data could be accessed
Quantifying and limiting	✓ Not only scans the sensitive data it looks for leakage too ✓ Could prevent against many attacks ✓ Could be effective used for all the three data states	× Doesn't ensure the complete protection × Limited to specific scenarios × Could disrupt the entire workflow process
Social and behaviour analysis	✓ Find out the malicious behaviour of users and detects data leaks ✓ Could be effective used for all the three data states	× Results in high False positive rates × Admin have to frequently access the system

Data identification	✓ More effective in unmodified data ✓ Low False positive Rates	× Data with more modification could not be detected × Lacking in data semantic understanding
Data Mining and Text Clustering	✓ Future data leaks could be predicted ✓ More effective in unclassified data ✓ More flexible and easily adaptable	× Need a learning phase × Results in higher False Positive Rates

VI. DATA ANALYSE PATTERN OF DLP

There are many analyse patterns by which the DLP analyse the data. They can be broadly classified in to two major group namely content analysis and context analysis. Content analysis mainly focuses on the primary data whereas context focuses on the surrounding of the data

A. Context analysis:

Various metadata properties of the confidential data are analysed. It scans the metadata information which is present along with the data and keeps a record of it[45]. The attributes include size, source, destination, data of creation, last modified, and similar properties, having a track of the details a process is created which provide detail policies used for data leak prevention.

B. Content analysis:

This analysis pattern focuses on the data itself instead of analysing the metadata of the sensitive data which may be text or any form of multimedia data[15]. Data leak is identified by comparing the transmission data with the existing data and if there is high similarity index found immediately it is blocked and also simultaneously reported to admin. The entire process of data loss prevention using three techniques namely, data fingerprinting, regular expression and statistical analysis.

VII. IMPLEMENTING DLP

Any sensitive data of any type of organisation could be prevented from leakage by effective
implementation of the DLP and they can be done by following some needed measures:

- ✓ Implementing a centralized DLP which takes care of all states of data.
- ✓ The right people with appropriate organization model has to be involved
- ✓ The data has to be rightly classified as sensitive and normal
- ✓ Implementation process has to be in phased manner
- ✓ The DLP implementation must not create any impacts in the workflow of the business process
- ✓ Reports must be more prominent
- ✓ Enough security measures must be taken before implementing the DLP

VIII. CONCLUSION

Data leakage is found to be a serious problem for any type of organization. In spite of the size of the organization data leaks can create a serious impact in their business and many proven examples has been found in the past, hence data leakage has to be prevented by any means. Most of the organizations have given more efforts in preventing their data from loss which might happen intentionally or accidentally. In finding the effective DLP both researchers and professionals are continually putting more efforts to make an effective DLP as DLPs are proved and recognised as the better solution for identifying, monitoring and protecting confidential data. The main aim of the survey was to find out the limitations of the existing DLPs, hence the security gaps could be identified and creating attention. In this work commercial DLPs have been listed along with their capabilities, also the academic DLPs also have been analysed their working method, advantages and limitations have also been discussed. The necessary implementation steps for an effective DLP has also been proposed.

From the survey it was observed that most of the existing DLPs suffer from critical limitations, which makes the DLP a compromising technology for the user. Most of the methods involve manual interpretation a mandatory one as they perform well in classified data. Hence the future DLP that has to be defined must not require manual influences for classifying the data instead artificial intelligence could be applied which might effectively classify the data and produce a better optimized DLP. The future DLP must be performing based on content based analysis instead of context based as content based could be more effective, as they retain a copy and compares the data. Finally the future system must be easy to manage and user friendly avoiding complexity and confusion.

REFERENCES:

1. "Risks, Threats, & Vulnerabilities in Moving to the Cloud," Carnegie Mellon University, 2018. [Online]. Available:https://insights.sei.cmu.edu/sei_blog/2018/03/12-risks-threats-vulnerabilities-in-moving-to-the-cloud.html.
2. B. Hauer, "Data and information leakage prevention within the scope of information security," IEEE Access, vol. 3, pp. 2554–2565, 2015.
3. B. Purohit and P. P. Singh, "Data leakage analysis on cloud computing," Int. J. Eng. Res. Appl., vol. 3, no. 3, pp. 1311–1316, 2013.
4. M. Prakash and G. Singaravel, "An approach for prevention of privacy breach and information leakage in sensitive data mining," Comput. Electr. Eng., vol. 45, pp. 134–140, 2015.
5. D. Sharing and D. Usage, "Privacy , Data Anonymization ," vol. 2002, pp. 5–11, 2012.
6. Ken Hess, "Reported data breached records in US from 2005 to present exceed 500 million | ZDNet," ZDNet, 2013. [Online]. Available: https://www.zdnet.com/article/reported-data-breached-records-in-us-from-2005-to-present-exceed-500-million/.
7. J.Clement, "U.S. data breaches and exposed records 2018 | Statista," Statista.com, 2019. [Online]. Available: https://www.statista.com/statistics/273550/data-breaches-recorded-in-the-united-states-by-number-of-breaches-and-records-exposed/.
8. M. Jain and S. K. Lenka, "A Review on Data Leakage Prevention using Image Steganography," vol. 5, no. 02, pp. 56–59, 2016.
9. Emma Bickerstaffe, "DATA LEAKAGE 1 What is data leakage prevention ?," Inf. Secur. Forum Ltd., vol. 4, pp. 1–18, 2018.
10. A. O. Karikari, "Detecting Data Leakage in Cloud Computing Environment," 2015.
11. S. Yoshihama and T. Matsumoto, "Web-based Data Leakage Prevention," Proc. Ithe 5th nternational Work. Secur. IWSEC 2010, 2010.

12. D. Pattanayak, “Effectiveness of Data Loss Prevention in Cloud Computing,” vol. 6, no. 2, pp. 364–368, 2016.
13. R. Nyarko, “Security of Big Data: Focus on Data Leakage Prevention (DLP),” 2018.
14. S. Alneyadi, E. Sithirasenan, and V. Muthukkumarasamy, “Detecting data semantic: A data leakage prevention approach,” Proc. - 14th IEEE Int. Conf. Trust. Secur. Priv. Comput. Commun. Trust. 2015, vol. 1, pp. 910–917, 2015.
15. B. Hauer, “Data and information leakage prevention within the scope of information security,” IEEE Access, vol. 3, pp. 2554–2565, 2015.
16. M. A. Haseeb, “MASTER ’ S THESIS Data Loss / Leakage Prevention,” 2013.
17. a R. P. Periyasamy and E. Thenmozhi, “Data Leakage Detection and Data Prevention Using Algorithm,” Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 7, no. 4, pp. 251–256, 2017.
18. A.Shabtai, “A Survey of Data Leakage Detection and Prevention Solutions,” vol. 2002, no. 2008, 2012.
19. S. Alneyadi, E. Sithirasenan, and V. Muthukkumarasamy, “Word N-gram based classification for data leakage prevention,” Proc. - 12th IEEE Int. Conf. Trust. Secur. Priv. Comput. Commun. Trust. 2013, pp. 578–585, 2013.
20. B. Hauer, “Data leakage prevention a position to state-of-The-art capabilities and remaining risk,” ICEIS 2014 - Proc. 16th Int. Conf. Enterp. Inf. Syst., vol. 2, pp. 361–367, 2014.
21. H. Guo, H. L. Viktor, and E. Paquet, “Identifying and preventing data leakage in multi-relational classification,” Proc. - IEEE Int. Conf. Data Mining, ICDM, no. December, pp. 458–465, 2010.
22. Ernst & Young, “Data loss prevention Keeping your sensitive data,” Insights governance, risk complianc, no. October, pp. 1–22, 2011.
23. M. Hart, P. Manadhata, and R. Johnson, “Text classification for data loss prevention,” HP Lab. Tech. Rep., no. 114, pp. 1–21, 2011.
24. M. Learning and F. Dlp, “Introduction to Forcepoint DLP Machine Learning How Forcepoint DLP machine learning works,” 2017.
25. Cloud Security Alliance, “SecaaS Implementation Guidance Category 2 // Data Loss Prevention,” Secaas Implement. Guid., no. September, 2012.
26. Frank Simorjay, “Data classification for cloud readiness,” Microsoft Trust. Comput., pp. 1 – 19, 2014.
27. D. Chen and H. Zhao, “Data security and privacy protection issues in cloud computing,” Proc. - 2012 Int. Conf. Comput. Sci. Electron. Eng. ICCSEE 2012, vol. 1, no. 973, pp. 647–651, 2012.
28. Y. J. Ong, M. Qiao, R. Routray, and R. Raphael, “Context-Aware Data Loss Prevention for Cloud Storage Services,” IEEE Int. Conf. Cloud Comput. CLOUD, vol. 2017-June, pp. 399–406, 2017.
29. M. Rouse, “What is data life cycle ? - Definition from WhatIs.com.” [Online]. Available: https://whatis.techtarget.com/definition/data-life-cycle.
30. M. Kazim and S. Y. Zhu, “A Survey on Security Threats in Cloud Computing Technology,” Int. J. Adv. Comput. Sci. Appl., vol. 6, no. 3, pp. 109–113, 2015.
31. G. Katz, Y. Elovici, and B. Shapira, “CoBAn: A context based model for data leakage prevention,” Inf. Sci. (Ny)., vol. 262, no. June 2002, pp. 137–158, 2014.
32. J. J. M and S. Manimurugan, “A Survey on Various Encryption Techniques,” Int. J. Soft Comput. Eng., no. 1, pp. 429–432, 2012.
33. R. I. Emori, “Scale models of automobile collisions with breakaway obstacles - Investigation indicates that scale models can be used to show the motion of breakaway signposts and lightposts after being struck by automobiles,” Exp. Mech., vol. 13, no. 2, pp. 64–69, 1973.
34. N. Mäkitalo, H. Peltola, J. Salo, and T. Turto, “VisualREST: A content management system for cloud computing environment,” Proc. - 37th EUROMICRO Conf. Softw. Eng. Adv. Appl. SEAA 2011, pp. 183–187, 2011.
35. R. Shaikh and M. Sasikumar, “Data classification for achieving security in cloud computing,” Procedia Comput. Sci., vol. 45, no. C, pp. 493–498, 2015.
36. H. Alhindi, I. Traore, and I. Woungang, “Data Loss Prevention Using Document Semantic Signature,” pp. 75–99, 2019.
37. P. Zilberman, S. Dolev, G. Katz, Y. Elovici, and A. Shabtai, “Analyzing group communication for preventing data leakage via email,” Proc. 2011 IEEE Int. Conf. Intell. Secur. Informatics, ISI 2011, pp. 37–41, 2011.
38. “Protecting Your Sensitive Data Everywhere.”
39. “Virtualization Security,” Virtualization Security, 2012. [Online]. Available: https://resources.infosecinstitute.com/virtualization-security-2/#gref.
40. X. Zhang, F. Liu, T. Chen, and H. Li, “Research and application of the transparent data encpryption in intranet data leakage prevention,” CIS 2009 - 2009 Int. Conf. Comput. Intell. Secur., vol. 2, pp. 376–379, 2009.
41. I. Journal, E. Technology, and C. Science, “PRESERVING SENSITIVE DATA BY DATA LEAKAGE PREVENTION USING ATTRIBUTE,” vol. 21, no. 3, pp. 705–711, 2016.
42. “Top 4 Reasons Why You Should Include Behavioral Analysis in DLP | Digital Guardian,” Digital Guardian, 2019. [Online]. Available: https://digitalguardian.com/resources/webinar/top-4-reasons-why-you-should-include-behavioral-analysis-dlp.
43. “SANS Institute: Reading Room - Intrusion Detection,” SANS, 2017. [Online]. Available: https://www.sans.org/reading-room/whitepapers/detection/paper/38165.
44. R. Jindal, R. Malhotra, and A. Jain, “Techniques for text classification: Literature review and current trends,” Webology, vol. 12, no. 2, pp. 1–28, 2015.
45. G. S. Chavan, S. Manjare, P. Hegde, and A. Sankhe, “A Survey of Various Machine Learning,” vol. 15, no. 6, pp. 288–292, 2014.