International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 12 6525 - 6530

Pushing the Limit of Data Leakage Protection Solution

Arzoo Arora
Department of Information Technology
K.J Somaiya College of Engineering
Mumbai, Maharashtra
arzooarora.arora7@gmail.com

Prof. Ravindra Divekar
Department of Information Technology
K.J Somaiya College of Engineering
Mumbai, Maharashtra
ravindradivekar@somaiya.edu

Abstract: Employees are the backbone of the organization, but they're also the biggest risk to the very data that makes the business thrive. Whether an insider is malicious in their attempt to take your confidential data for personal gain, or they just don't know better and mishandle confidential data thereby putting it at risk, insiders significantly contribute to data loss. The sensitive data could be customer information, personal details, intellectual property and many more. "Data Loss" and "Data Leak" are often used interchangeably. To keep corporate data safe, people, processes and technology must holistically address the insider threat. Different companies offer the data loss prevention (DLP) solution to protect data at rest, in motion and in use.

Keywords- Data Loss Prevention, Endpoint DLP, Network DLP, Storage DLP, Exact Data Match

*****

I. INTRODUCTION

Before stating the requirement of DLP, it is important to understand what exactly the DLP is and when it hit the market first. The term DLP, which stands for Data Loss Prevention, first hit the market in 2006 and gained some popularity in early part of 2007 [1]. Just as we have witnessed the growth of firewalls, intrusion detection systems (IDS) and numerous security products, DLP has already improved considerably and is beginning to influence the security industry [1]. While DLP has been known by several acronyms, in simple terms, it is truly a technology that provides visibility at content level into one's network [1]. It is basically designed to detect potential data breach / data ex-filtration transmissions and prevent them by monitoring, detecting and blocking sensitive data while in-use (endpoint actions), in-motion (network traffic), and at-rest (data storage) [2]. So it helps to ensure that the sensitive data is not going to the unauthorized persons and also the monitoring is performed for confidential data in the network. In case of data breach and data exfiltration, sensitive data is disclosed to unauthorized personnel either by malicious intent or unknowingly.

II. NEED FOR DATA LOSS PREVENTION

Data is more accessible and transferable today than ever before, and the vast majority of data is sensitive at various levels [3]. Some is confidential simply because it is part of an internal organization and was not meant to be available to the public. Some data is sensitive because of corporate requirements, national laws, and international regulations. Often the value of data is dependent upon its remaining confidential - consider intellectual property and competition [3].

Leakage of your data could be embarrassing or worse, cost you industrial edge or loss of accounts. Allowing your organization to act in non-compliance with privacy acts and other laws could be worse than embarrassing - the integrity of your organization may be at stake [3].

Most of the time, the leakage of data happens unintentionally. The best solution to prevent unintentional data leaks is to implement an automated corporate policy that will catch protected data before it leaves your organization. Such a solution is known as Data Loss Prevention (DLP).

Data Loss Prevention identifies, monitors, and protects data transfer through deep content inspection and analysis of transaction parameters (such as source, destination, data object, and protocol), with a centralized management framework [3]. In short, DLP helps in detection and prevention of the unauthorized transmission of confidential information.

III. DETECTION TECHNOLOGIES

DLP can accurately detect all of the confidential data in your organization—whether it's at rest, in motion, or in use. The detection technologies include:

A. ***Exact Data Matching (EDM)*** detects content by fingerprinting structured data sources, including databases, directory servers, or other structured data files [6].In this, fingerprinting of sample data has been performed for exact data match.

B. ***Indexed Document Matching (IDM)*** applies fingerprinting methods to detect confidential data stored in unstructured data, including Microsoft Office documents; PDFs; and binary files such as JPEGs, CAD designs, and multimedia files. IDM also detects "derived" content, such as text that has been copied from a source document to another file [6].Here, you can prepare a policy on the basis of different extensions.

C. ***Vector Machine Learning (VML)*** protects intellectual property that has precise characteristics that may be rare or difficult to describe, such as financial reports and source code. It detects this type of content by performing statistical analysis on unstructured data and comparing it to similar content or documents present in structured form. Unlike other detection technologies, VML does not require you to locate, describe, or fingerprint the data you need to protect [6].

 ISSN: 2321-8169
6525 - 6530

D. ***Described Content Matching (DCM)*** detects content by looking for matches on specific keywords, regular expressions or patterns, and file properties [6]. Most of the DLP provides more than 30 data identifiers out-of-the-box, which are pre-defined algorithms that combine pattern matching with built-in intelligence to prevent false positives.

E. ***File type detection*** recognizes and detects more than 330 different file types such as email, graphics, and encapsulated formats. You can configure the DLP to recognize virtually any custom file type, and it also allows you to extract content from specific file formats—including encrypted formats—using the Content Extraction API [6].

IV. DLP CATEGORIES

DLP solution is designed to achieve three key objectives. These key objectives help to protect organization's most valuable assets and keeping them out of the public domain. On the basis of above mentioned objectives, we can make three different categories i.e. Storage DLP (data at rest), Network DLP (data in motion) and Endpoint DLP (data in use) which are explained below:

A. ***Storage DLP (Data at Rest):*** The basic function of DLP solutions is the ability to identify and log where specific types of information are stored throughout the enterprise. This means that the DLP solution has the ability seek out and identify specific file types - such as word documents, excel sheets and many other whether they are on file servers, storage area networks (SANs) or even endpoint systems. This discovery action is performed on the basis of policies defined by the person handling the DLP in the enterprise. . Once found, the DLP solution is able to open these files and scan their content to determine whether specific piece of information are present, such as customer personal information or any sensitive data. risk associated with this type of data include the lack of visibility into where sensitive data is stored, the lack of understanding around who has access to the sensitive data, and the lack of secure storage for sensitive data to prevent theft and loss.

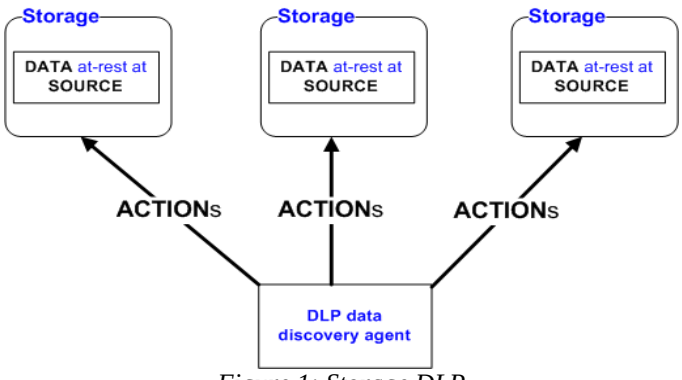

Figure 1: Storage DLP

DLP solution helps to convert the unstructured data to structured one via identifying the presence of sensitive data at some place and the owner of that data

B. ***Network DLP (Data in Motion):*** Data in motion consists of information that is electronically transmitted outside an organization's network via email, online chat rooms and other methods. Common risk associated with this type of data include the loss of sensitive data through various communication mediums, the harvesting of sensitive data by malware and broken business processes that expose sensitive data. DLP solution could stop the sensitive data loss through electronic mails and web. It does not just monitor what data is going out of the enterprise, but also gives us an option to either monitor or block its content after checking the data with respect to the policy. It is possible to do the fingerprinting of data with the sender and the recipient also which is a more effective way to protect the sensitive data.

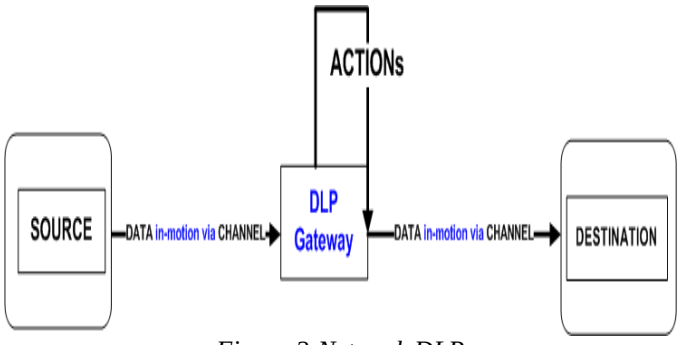

Figure 2:Network DLP

C. ***Endpoint DLP (Data in Use):*** Data at the endpoint relates to the information stored on laptops and portable storage devices. Data at end-points is an agent based solution that sits on end user workstations and laptops monitoring for any data leaving via removable devices, such as floppies, CDs, USBs, etc. DLP products help protect sensitive data even when equipment is offline by identifying sensitive data stored on portable storage devices and restricting use of those devices. This also provides auditing and protection against users printing classified data. End-user notification and self-remediation options notify users of policy violations while allowing them to continue business in compliance with corporate data-handling policies designed to DLP solution is designed to address three different scenarios. Below are the three different categories which are formed on the basis of three different scenarios

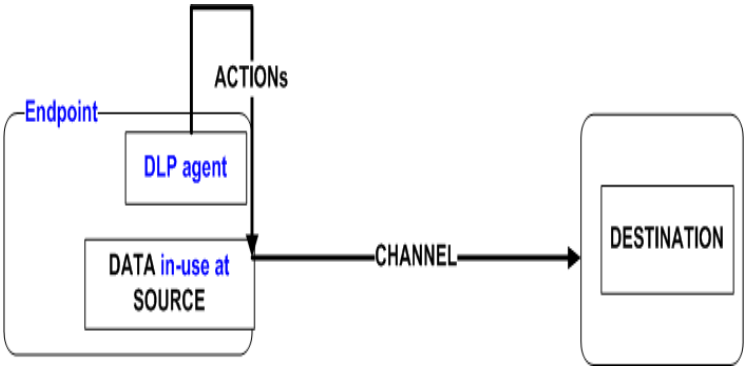

Figure 3:Endpoint DLP

V. DLP CATEGORIES

The potential legal liability and damage to brand-reputation from exposure of sensitive data has encouraged security leaders to implement a DLP program to prevent both the intentional and unintentional exposures. A DLP program can be constructed in many ways. Most DLP solutions solely rely on technology. Although technology is an important aspect, it also takes people and process to build a holistic DLP program.

The following diagram illustrates a unique model:

A. ***Data Governance:*** Data Governance (DG) encompasses the overall management of the basic components of data protection i.e. confidentiality, integrity and availability of data within an enterprise. The detailed components of DG can be difficult to label because it is a new concept and it is still evolving [7]. From a security standpoint, DG is the act of protecting data and monitoring the flow of where the data travels [7]. Although this may sound overly simplified, this can be a challenging task within a large organization [7]. A sound DG program includes a governing body or a committee to define policies and procedures and a plan to implement those procedures [7]. In a large organization, this group needs to consist of individuals who have a strong understanding of the organization's industry, business objectives, internal processes, and the corporate culture [7]. The group will be responsible for the creation and implementation of the policies and rules as per the department requirements for the prevention of sensitive data. The policy maker has to define the response against all the policies implemented in production. Before getting it live, they usually checks the working of the policy in UAT environment.

B. ***Risk Assessment:*** Conducting a risk assessment is the most important first step in any DLP program. The main purpose for a Risk Assessment is to identify all types of data both structured and unstructured within your network and to identify threats and vulnerabilities related to this data and the possible ways to protect it. Once this information has been identified, a flow analysis needs to be conducted to identify all systems and devices the data either resides on or flows through by checking the data lifecycle. For example, the HR department may utilize employee personal and salary related information. This information is stored on a centralized server utilizing a second server with a proprietary database [7]. The HR employee connects their intranet web browser to the server (i.e., three-tier architecture) to fetch the data. In this simple scenario, the devices transferring and storing data are the employee's desktop workstation, network components connecting to the server, the server itself, and a separate server maintaining the proprietary database [7]. Each of these systems needs to be evaluated to determine threats and vulnerabilities that may put the data at risk, so that proper recommendation could be implemented. This exercise needs to be conducted for all types of data being utilized within the organization from security point of view.

C. ***Regulatory and Privacy Requirements*****:** One key step in a DLP program is to identify regulatory requirements [7]. Having a strong understanding of what regulatory requirements apply to your organization and what types of security controls are required, need to be identified to have an appropriate policy. Most organizations do not have a strong understanding of their requirements, or their interpretations of those requirements are different from the regulators [7]. Thus, most organizations are operating in a non-compliance mode without knowing the reality. Identifying regulatory requirements supports the prioritization security resources-to-system containing, processing, and/or transmitting the sensitive information [7]. This also helps to focus the scope of security controls in the organization. Identifying privacy requirements is essential for an organization to ensure that the goals and promises of privacy and confidentiality are supported by its practices, thereby protecting confidential information from abuse and the organization from liability and public relations problems [7]. Although there are some federal and state regulatory requirements, most organizations maintain privacy policies and procedures to satisfy the comfort levels of their customers [7]. A successful DLP program needs to conduct a privacy assessment to ensure that data is protected based on the organization's policies. Some DLP's also provide default templates on the basis of regulatory requirements.

Figure 4: DLP Components

D. ***Data Classification:*** Data Classification is the process of classifying information data according to its value and sensitivity to the organization as it varies from one organization to another on the basis of their type of work. Data classification provides the proper prioritization of an organization's assets and resources, which will result in the appropriate level of policies be designed to each system accordingly. Data should be categorized in terms of criticality within an organization's environment (i.e., public, confidential, secret, and private, etc.). Once the data has been identified,classification categories can be assigned appropriately.

E. ***Policies, Standards, Procedures:*** Sound policies, standards, and procedures are fundamentals for a DLP to work effectively. They ensure that the data is protected at a level appropriate to its organization's value. It is critical not only to create sound policies, standards and procedures, but also to ensure that they are updated on a regular basis and are specifically for each department. Within the realm of DLP, policies are the starting point before a company can establish standards and procedures, which allow an organization's DLP solution to operate more securely and efficiently [7]. Standards are mandatory activities, actions, rules, and regulations designed to provide the DLP policies with the support, structure, and specific direction required to be meaningful and effective [7]. Procedures spell out the specifics of how the DLP policies and the supportive standards will actually be implemented in an operating environment [7].

F. ***Data Discovery:*** Regardless of the amount of security controls implemented, the chances of intellectual property leaking out are highly likely because a lot of data is present in an unstructured form and sometime employees are not aware of the sensitivity of the data they have. This is why a data discovery assessment needs to be conducted on a regular basis. Data discovery is one of the key elements of a DLP program. Access to a strong discovery tool and knowledgeable staff can limit most organizations from implementing a solid DLP program.

G. ***Remediation Processes:*** A major challenge with DLP programs is determining which data is valid leaked data and which data is a false positive detection in the console. In today's business environment, the amount of data traveling through the network and stored on disk drives is almost overwhelming. Nevertheless, the challenge needs to be managed and processes need to be in place beforehand. Once the incident has been triggered against any policy in the console, team has the defined responses in the policy that automatically goes to the business owner. After a violation has occurred, an investigation needs to be launched to determine if a corporate policy has been violated. To accomplish this objective without disrupting the work environment, processes and procedures need to be in place to effectively remediate the issue [7]. A strong resolution process needs to be automated, efficient, and timely to manage and resolve the issue before the organization is harmed [7].

H. ***Training and Awareness:*** It is important for an effective DLP solution to interact with the organization's employees so that they have a strong understanding about what type of data they are having, why certain activities are inappropriate and could be harmful for the organization. Not all violations are conducted with a harmful intent during work. An employee may want to work at home and email sensitive data to their personal, less secure public account without knowing the consequences. Although the intent is good, the action is not. Ongoing education will help reinforce correct behavior and provide the employee with guidance on which data is sensitive and how to correctly handle it.

VI. DLP ARCHITECTURE

A. ***Enforce Platform:*** In short, we can say Enforce platform is one where policies can be defined and administration can be done [8]. Incidents can also be reviewed in the console. Enforce Platform is the central web-based management console and incident repository [8]. It is where we define, deploy and enforce data loss policies, respond to incidents, analyze and report policy violations and perform system administration of the incidents. Enforce Platform enables us to write policies once and also allows fine tuning. It helps to monitor the incidents triggered on the basis of defined policies in the console. It uses three classes of detection technologies to provide complete and accurate coverage across your endpoint, network, and storage systems:

1) Describing protects structured and unstructured data by looking for content matches on keywords, expressions or patterns, and signatures [8].

2) Fingerprinting protects structured and unstructured data by looking for exact or partial content matches on indexed data sources and documents [8].

3) Learning protects unstructured textual data by building a statistical model using example documents and calculating content similarity [8].

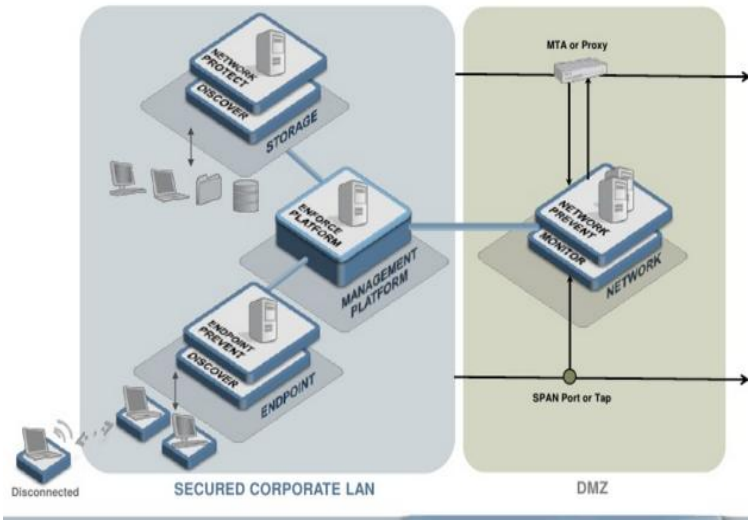

Figure 5:DLP Architecture

B. ***Endpoint Server:*** DLP Agent is necessary component of Endpoint Channel. The DLP Endpoint Agent provides control of Data Loss Prevention policies and manage the data on those machines. The DLP Endpoint Agent is made up of two agents, the endpoint agent and the watchdog agent. These two agents watch each other to make sure they are still running and will restart the service If one of those services are started. With the endpoint agent, policies applied to the Data at Rest targets and the network via Data in Motion can be applied to laptops and desktops. All scans on endpoints are controlled through the agent and information is reported to the Enforce server.

C. ***Database Server:*** Database Server is very important part of Data Loss Prevention System which mainly stores each and every incident triggered on the basis of policies. We can also fetch the reports through console from the database for incidents.

D. ***Network Prevent:*** Network Monitor passively inspects network traffic for confidential data that is being sent in violation of security policy. It is deployed at egress points in the network demilitarized zone (DMZ) with a network tap or Switched Port Analyzer (SPAN) [9]. Network Monitor performs deep content inspection of all network communications with no packet loss, unlike other data loss prevention (DLP) solutions which may only sample packets during peak loads and put you at greater risk for data loss [9].

1) Network Prevent for Email redirects, quarantines, or blocks outbound messages containing sensitive data as per the policy. It is deployed at egress points in the network DMZ and integrates with your existing on-premise, public or private cloud messaging infrastructure [9]. Network Prevent for Email provides comprehensive protection for managed and unmanaged endpoints, including mobile devices that access your corporate email systems [9].

2) Network Prevent for Web blocks or removes sensitive data from outbound web communications if they violate security policy by uploading the data to an unauthorized portal. It is deployed at egress points in the network DMZ and integrates with your on-premise, public or private cloud messaging infrastructure which is same in both the email and web case. Network Prevent for Web provides comprehensive protection for managed and unmanaged endpoints, including mobile devices, that access the Web through your corporate network [9].

E. ***Discover Server:*** Discover Server locates a wide range of exposed confidential data in the organization. It communicates with the Enforce Server to obtain information about policies and scan targets. It sends information about the exposed confidential data that it finds to the Enforce Server for reporting and remediation which helps to monitor the data. Sensitive or at-risk data can include credit card numbers or names, addresses, and identification numbers. Endpoint Discover examines the local fixed drives of an endpoint and locates every file that contains the information that matches a policy. Endpoint Discover scans the endpoints to find the information that you have defined as at risk or sensitive.

VII. DATA SECURITY MECHANISM

A. ***Identity Finder DLP Console*** - The DLP Console uses multiple layers of security when processing and storing sensitive information received from DLP Endpoints, entered locally, and when publishing sensitive configuration data for DLP Endpoints.

1) The SQL Database Server employs password based encryption using a passphrase entered during initial installation. This passphrase can be changed at after installation. The encryption is performed internally by SQL Database Server using the Transact-SQL functions ENCRYPTBYPASSPHRASE and DECRYPTBYPASSPHRASE. All identity match data, sensitive DLP Console configuration data, and sensitive configuration data used to manage DLP Endpoints is encrypted in this manner. The database encryption passphrase is stored in an encrypted state on the IIS Web Server and can only be retrieved by components of the DLP Console suite.

2) During installation of the DLP Console, an RSA 1024 bit public/private encryption key pair is generated. The private key is stored only in the SQL Database Server and can only be obtained by the IIS Web Server using the database encryption passphrase. These keys are used to encrypt information stored and subsequently transmitted by the DLP Endpoints to the DLP Console.

3) The IIS Web Server can be configured to use HTTP, HTTPS, or both HTTP and HTTPS on any port. By default HTTP will run on port 80 and HTTPS on port 443 and may be changed. Because all sensitive data is encrypted prior to being transmitted to the DLP Services Web Application, it is not necessary to enable HTTPS ; however HTTPS may be enabled to provide additional security. The DLP Console Web Application can be configured to suppress the display of any sensitive data or it can be configured to display everything. To ensure encryption of data between a web browser and the DLP Console Web Application, HTTPS should be used.

B. ***Identity Finder DLP Endpoints*** – The DLP Endpoints also utilize multiple security mechanisms to ensure that sensitive data is encrypted.

1) Installation of the DLP Endpoint requires configuration information including the location URL of the DLP Services Web Application and the public key from the encryption key pair generated by the DLP Console installation.

2) All sensitive data sent to the DLP Console is encrypted in a SQLite database using, by default, 128 bit AES encryption. Alternatively, this can be configured to use 256 bit AES encryption or RC-4 encryption. The passphrase used to encrypt the SQLite database is randomly generated and transmitted to the DLP Console. The passphrase is encrypted using the 1024 bit RSA public key from the public/private key pair generated by the DLP Console. If the public key is not available or has been removed from the DLP Endpoint, a strong internal password known only to the DLP Endpoint and DLP Console software is used to encrypt sensitive data.

3) When saving encrypted results files, the DLP Endpoint uses AES 256 bit FIPS 140-2 validated encryption.

VIII. CURRENT CHALLENGES AND THEIR SOLUTIONS

After understanding the Data Leakage Protection Tool in general, I have studied about the specific DLP solution i.e. Symantec DLP. I would like to add the limitations of the existing DLP system and their possible solutions also. By considering these enhancements, we can increase the level of data protection through DLP Tool. The possible enhancements

are given below with respect to every limitation which will help us to have a better DLP solution:

A. First limitation is that DLP solution is, it's un-ability to understand the Encrypted Password Protected Zip files. When these kind of files crosses the DLP, then it does not get what to do with the file. Because when it tries to open the file then it will ask a password and dlp does not get any way to open the file and that's why allows it to pass the DLP. File could contain any sensitive data of the organization whose leakage could impact the Organization's good will. Solution to have a look into these type of files is, DLP should be able to understand that it is a Encrypted Password Protected file. But the binary values of all these type of files are different, so it is necessary to add this feature in the dlp where it could recognize the file from the dialogue box for the password.

B. Nowadays, SFTP Protocol is widely used in the organizations to share the data with the vendor. It works on the SSH traffic and DLP is unable to monitor the SSH traffic. DLP could only monitor the SSL traffic. So it is required to add a functionality in the DLP for the monitoring of the SSH traffic.

C. DLP agents are normally installed on the systems for the endpoint protection. Any data leaving via removable devices, such as floppies, CDs, USBs, etc. DLP helps to protect sensitive data even when equipment is offline by identifying sensitive data stored on portable storage devices and restricting use of those devices. User can also uninstall the DLP Agent, but for the un installation a password is required. If anyhow a password get compromised, then we will never come to know in the console about this. Solution for this could be a notification that should trigger in the console like an incident if somebody does the un installation.

D. In organizations, DLP is usually integrated with the AD's which ensures that DLP could monitor each and every asset present in the organization. Normally policies are developed on the basis of asset code of the user's system or on the basis of email ID's. But we cannot make a policy for an AD Group, as it does not allow us to do so. Solution for this problem is that it should have a feature where we can create a policy on the basis of different groups in the AD.

E. Policy definition depends upon the business requirement for data protection. When creating a content type rule, there is a very useful option: Count all matches and only report incidents with atleast matches. But when creating a file properties type rule, there are no such options. For example, we can create a "Message attachment or File Type Match" rule to find specific type of documents such as office and pdf files. It will be great to add matches count option in this rule. If so, the user can add a rule to detect an incident that contain much more 3 matches with specific type of document. It helps to make a more refined policy in the DLP.

F. Enhanced Application monitoring in DLP: When it comes to DLP Agents, we can enable or disable Application Monitoring in the agent configuration. When application monitoring is enabled, all the applications listed in the application monitoring list with their respective selected channels are monitored. The only way to disable monitoring any pre-defined application is to deselect all the channels for that application. It would be better if we can enable/disable entire applications in the application monitoring list so that the we do not disturb the original configuration. Moreover it would be great if we can have multiple application monitoring lists each allowing defining its own applications and enabling/disabling pre-defined applications. Such lists can then be assigned to the agent configuration. Such flexibility will help in having multiple application monitoring configurations based on requirements.

G. It seems to be that we are finding many customers who are looking to use different policies with different applications. One example: For SFTP applications, they want to present users with a User Cancel prompt as a response for the incident, while for Cloud File Syncing apps (dropbox, box.net, etc) they want to completely block the application. It seems many customer would like to have the option to turn on certain applications to begin with, and secondly, would like to have control to assign certain applications (listed in the Application File Access Control section) to a group that can be specified in policy rules and/or response rules.

IX. CONCLUSION

Nowadays, companies deal with the sensitive data which is the main area of concern. So there should be a solution which can help us out for the data protection. This paper provides an idea about what do you mean by data protection solutions and why it is required in an organization. It is basically designed to detect potential data breach / data ex-filtration transmissions and prevent them by monitoring, detecting and blocking sensitive data while in-use (endpoint actions), in-motion (network traffic), and at-rest (data storage). With the above mentioned enhancements, we can make more requirement specific policies and also will be able to do the blocking and monitoring in a better way by pushing its limits..

REFERENCES

[1] https://www.sans.org/reading-room/whitepapers/dlp/data-loss-prevention-32883
[2] https://en.wikipedia.org/wiki/Data_loss_prevention_software
[3] https://sc1.checkpoint.com/documents/R77/CP_R77_DataLossPrevention_AdminGuide/62453.htm
[4] http://blog.agilis.com.tr/symantec/dlp/en/2013/04/24/symantec-data-loss-prevention-new-features-protect-against-insider-threats
[5] http://www.isaca.org/Groups/Professional-English/security-trend/GroupDocuments/DLP-WP-14Sept2010-Research.pdf
[6] https://www.symantec.com/content/en/us/enterprise/fact_sheets/data-loss-prevention-solution-ds-21350666.pdf
[7] http://www.mcafee.com/in/resources/white-papers/foundstone/wp-data-loss-prevention-program.pdf
[8] http://www.symantec.com/content/en/us/enterprise/fact_sheets/b-dlp_enforce_platform_DS_21189690.en-us.pdf
[9] Symantec™ Data Loss Prevention Administration Guide