Data leakage analysis on cloud computing

[1]Bijayalaxmi Purohit, [2]Pawan Prakash Singh
[1, 2] (Department of Computer Science Engineering, Suresh Gyan Vihar University, Jaipur

ABSTRACT

Cloud describes the use of a collection of services, applications, information, and infrastructure. It is like a pool of resources and services available in a pay as- you-go manner. Services like computation, network, and information storage. This paper mainly focus on the major security concerns about cloud computing. The major areas of focus are: - Information Protection, Virtual Desktop Security, Network Security, and Virtual Security. In today's business world, many organizations use Information Systems to manage their sensitive and business critical information. The need to protect such a key component of the organization cannot be over emphasized. Data Loss/Leakage Prevention has been found to be one of the effective ways of preventing Data Loss. DLP solutions detect and prevent unauthorized attempts to copy or send sensitive data, both intentionally or/and unintentionally, without authorization, by people who are authorized to access the sensitive information. DLP is designed to detect potential data breach incidents in timely manner and this happens by monitoring data.

Keywords **–** Data leakage analysis in cloud computing; Data leakage prevention in cloud computing; Checking sensitivity of data; Data security in cloud

I. INTRODUCTION

Data loss, which means a loss of data that occur on any device that stores data. It is a problem for anyone that uses a computer. Data loss happens when data may be physically or logically removed from the organization either intentionally or unintentionally. The data loss has become a biggest problem in organization today where the organizations are in responsibility to overcome this problem. Data Leakage is an incident when the confidentiality of information has been compromised. It refers to an unauthorized transmission of data from within an organization to an external destination. The data that is leaked out can either be private in nature and are deemed confidential whereas Data Loss is loss of data due to deletion, system crash etc. Totally both the term can be referred as data breach, has been one of the biggest fears that organization face today.

Data Loss/Leakage Prevention (DLP) is a computer security term which is used to identify, monitor, and protect data in use, data in motion, and data at rest[1]. DLP is sued to identify sensitive content by using deep content analysis to per inside files and with the use if network communications. DLP is mainly designed to protect information assets in minimal interference in business processes. It also enforces protective controls to prevent unwanted incidents. DLP can also be used to reduce risk and to improve data management practices and even lower compliance cost. Systems are designed to detect and prevent unauthorized use and transmission of confidential information. Vendors refer to the term as Data Leak Prevention, Information Leak Detection and Prevention (ILDP), Information Leak Prevention (ILP), Content Monitoring and Filtering (CMF), Information Protection and Control (IPC) or Extrusion Prevention System by analogy to Intrusion-prevention system [1].

In this paper, we deal with both the terms data loss and data leakage in analyzing how the DLP technology helps in minimizing the data loss/leakage problem? The study is performed as a case research on DLP technology in organizational perspective.

II. BACKGROUND

Research Approach

There are three kinds of research approaches in scientific research; Quantitative research, Qualitative research and Mixed research. Different researcher gives different definitions to qualitative research, qualitative research and mixed research. Here are some of them; Quantitative Approach: A quantitative research is the one which involves strategies of inquiry such as experiments and surveys, and collects data on predetermined instruments that yield statistical data. Generally, the quantitative research aims at explanation which answers primarily to why? Quantitative data collection is based on precise measurement using structured and validated data collection instruments such as closed ended items, behavioral responses and rating scales.

In addition, quantitative research is defined as social research that employs empirical methods and empirical statements. The author states that an empirical statement is defined as descriptive statement about "what is the case in the real world" rather than "what ought to be the case"[2].

Therefore quantitative research is essentially about collecting numerical data to explain a particular phenomenon, particular questions seem immediately suited to being answered using quantitative methods. Qualitative Approach: A qualitative research is the one which involves strategies of inquiry such as narratives, phenomenology's, ethnographies, grounded theory studies, or case studies.

Generally, the qualitative research is a type of scientific research that aims at understanding which answers primarily to how? Qualitative data collection is based on in-depth interviews, participant observation, field notes and open-ended questions. Here the research is the primary data collection instrument.

"Participant observation [for collecting data on naturally occurring behavior's in their usual contexts], In-depth interviews [for collecting data on individual perspectives, and experiences], and Focus groups [also called as group interviews is effective on eliciting data on the cultural groups]" are some kinds of qualitative research methods[3].

Mixed Approach: A mixed research involves the mixing of quantitative and qualitative methods. The mixed approach involves strategies of inquiry such as collecting data either simultaneously or sequentially to best understand research problem. The data collection involves gathering both numeric information as well as textual information. The study begins with a broad survey and then focuses on qualitative, open-ended interviews to collect detailed views from participant. There are three ways of mixing the data's such as merging the data, connecting the data, and embedding the data. Though it is not enough to simply collect and analyze the data's (both quantitative and qualitative) there is a need to be mixed together in order to form a complete picture of the problem then they do when standing alone. From the above details, we then believe our research is of qualitative approach. Therefore the research needs not to know statistical analysis as the quantitative approach suggest. The need to conduct this research is to know the detailed understanding of how the DLP technology minimizes the data loss problem in the organization[4].

Research strategy

Generally, research strategy is a way of collecting and analyzing empirical evidence by following some logic. A research design is the logic that links the data to be collected and the conclusions to be drawn to the initial questions of the study, it ensures coherence. There are five major research strategies; experiments, research survey, archival analysis, histories, and case studies. Each strategy has its own strength and weakness and can be utilized for all three research purposes; exploratory, descriptive, and explanatory. Case study research involves the study of an issue explored through one or more cases through a boundary system. The author also states that it is a qualitative approach in which the investigator explores a case in detailed, and in depth data collection involving multiple sources of information and depicted a case description and case based themes[5]. The intent of case analysis exists in three variations such as single instrumental case study, the multiple case studies, and the intrinsic case study[5]. In a single instrumental case study, and then selects one bounded case to illustrate the issue. In a collective case study, the one issue is again selected but the inquirer selects multiple case studies to illustrate the issue. The intrinsic case study focuses on the case itself because the case presents an unusual and unique situation.

This research therefore is designed in form of a case study, a single instrumental case study to be more definite. The research focuses on phenomenon, which is "How do the DLP technology helps in minimizing the data loss/leakage problem in conjunction with previously used technologies in the organization?" and it was examined in there different ways such as the product, people and process[6].

Data Collection Method

Generally, qualitative research often emphasizes the human factor to understand their behavior, knowledge, altitudes and fears. The qualitative research involves qualitative data that are obtained through methods such as surveys or interviews, on-site observations, and focus groups. Data are the empirical evidence or information one gathers carefully according to rules or procedures‟ [7]. We also found that aim of data collection strategy is to obtain answers from different sources and this will let the researcher to describe, compare, and relate one characteristic to another and demonstrate that certain feature exist in certain categories.

Case study is a qualitative approach in which the investigator explores a case in detailed, and in depth data collection involving multiple sources of information (such as observation, interviews, documents, audio visual materials) and reports a case description and case based themes.
Basically, there are two types of data collection methods; Primary and secondary[7]. Primary data collection: This processes three different types of strategies; interview, questioning, and observation. It is the most substantial method in all qualitative inquiry. It is first-hand information collected through various methods such as observation, interviewing, mailing, etc.

Secondary data collection: This has been collected and processed by other researchers for different purposes than what it is sued for. It is a very common practice to collect, process, utilizes, and store data by companies and organizations for the support of their operation. The secondary data are mostly collected from sources such as magazine, news paper, TV, internet, reviews, and research articles.

For this research, interviews, observations, documents, and reports have been extensively used as a form of data collection. Main data‟ s are captured from the company internal knowledge base (real time data or empirical data) as one of our researcher is working for the organization on Data Loss Prevention project and another researcher have worked in conducting interview questions in to gather the project details with respect to thesis outline. Both closed and open ended questions were used during the interviews, and the interviews were performed in email system. Along with this, security journals, DLP books such as (Data Leak Prevention - ISACA), and are used in collecting the data.

Classification of Information Leakage

From the paper, the author classified the information leakage into three levels which means a document containing confidential data can be classified as unintentional leak, intentional leak, and malicious leak [8].

Unintentional Leak:

1. Attach document
2. Zip and send
3. Copy & Paste

The unintentional leakage normally occurs when a user mistakenly sends a confidential data or information to third party or wrong recipient. This is done without any personal intention. For instance, if an employee sends an email attaching a document mistakenly this contains confidential data to a wrong person or to vendor.

Intentional Leak:

The intentional leakage normally occurs when a user tries to send a confidential document without aware of company policy and finally sends anyhow. This is usually done when a user bypassing the security rules and regulations or devices without trying to gain personal benefits. For instance, when an employee renames a document folder and partially copies the data from it.

Intentional Leak

1. Document renames
2. Document type change
3. Partial data copy
4. Remove keyword

Malicious Leak:

Malicious leakage usually caused when a user deliberately trying to sneak the confidential data past the security rules

Malicious Leak

1. Character encoding
2. Print screen
3. Password protected
4. Self extracted archive
5. Hide data
6. Policies or product.

For instance, when an employee sneaks a confidential data from enterprise system and sends them through email and even cause vulnerability to the system.

III. DATA LEAKAGE PREVENTION USING MY DLP

MyDLP is open source all-in-one data loss prevention software that runs with multi-site configurations on network servers and endpoint computers. MyDLP development project has made its source code available under the terms of the GNU General Public License.

MyDLP is one of the first free software projects for data loss prevention.

MyDLP allows you to monitor, inspect and prevent all outgoing confidential data without the hassle. With painless deployment and configuration, easy to use policy interface and great performance IT administrators and security officers are able to combat data leakage.

With MyDLP you can;

1. Block or quarantine outgoing confidential data from your organization network via mail and web. Archive suspicious files.
2. Monitor removable device usage in your organization and block or quarantine confidential files copied into these devices such as USB memory sticks or smart phones.
3. Block or quarantine print jobs which contain confidential information.
4. Discover confidential data on network storages, databases, workstations and laptops in your organization.

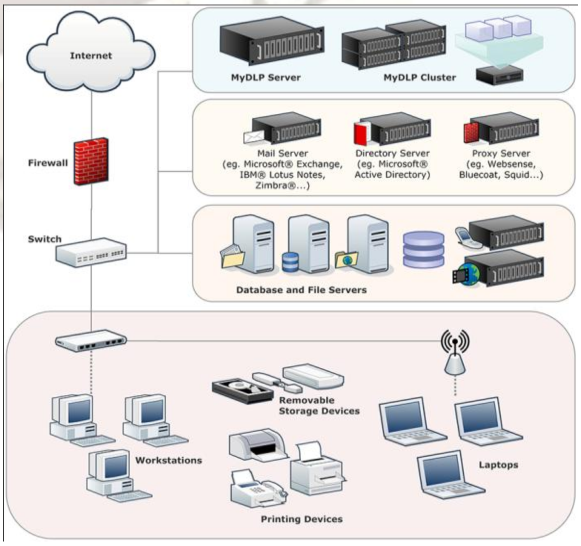
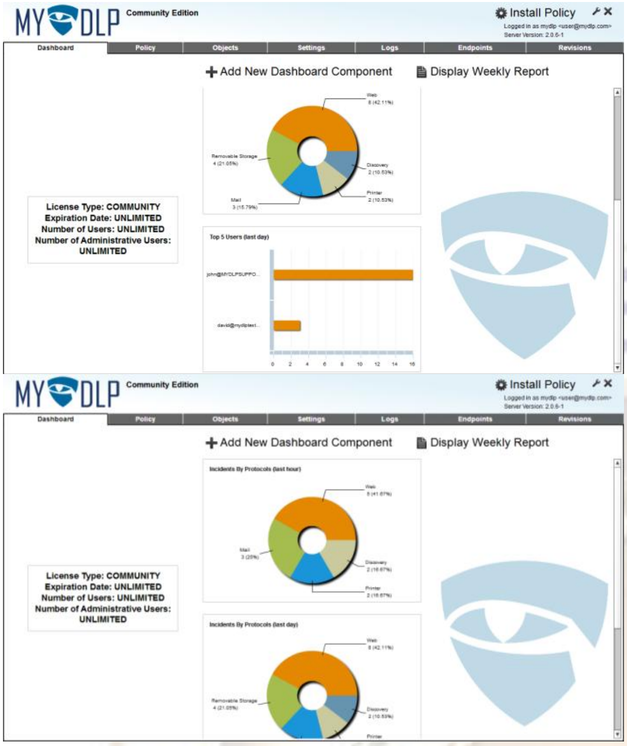

Internet
MyDLP Server
MyDLP Cluster
Firewall
Mail Server (eg. Microsoft® Exchange, IBM® Lotus Notes, Zimbra®...)
Directory Server (eg. Microsoft® Active Directory)
Proxy Server (eg. Websense, Bluecoat, Squid...)
Switch
Database and File Servers
Removable Storage Devices
Workstations
Laptops
Printing Devices

Figure1. End Points in cloud computing.

I. EXPERIMENT AND RESULT



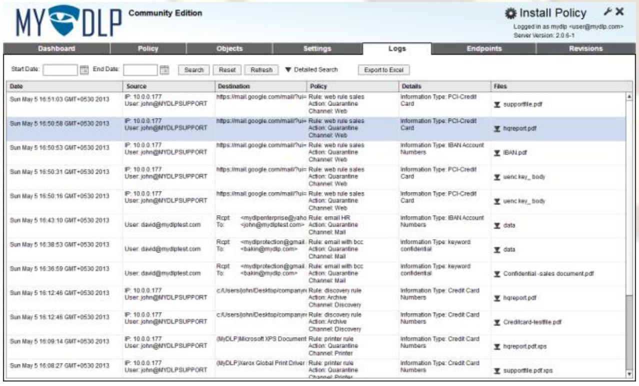
License Type: COMMUNITY
Expiration Date: UNLIMITED
Number of Users: UNLIMITED
Number of Administrative Users: UNLIMITED

License Type: COMMUNITY
Expiration Date: UNLIMITED
Number of Users: UNLIMITED
Number of Administrative Users: UNLIMITED

Figure2. Dash board of My DLP.

Below is the excerpt of the dash board of DLP suite deployed. The dash board is a graphical user interface which provides complete snapshot of the DLP. It categorizes incidents by Network, Endpoint and Datacenter. It also shows number of events or incidents generated and their status.

Based on both the type of policy and content blade, number of incident triggered can be seen in the dashboard. It's very user friendly environment with lots of information on a mouse click.



Date	Source	Destination	Policy	Details	Files
Sun May 5 16:51:03 GMT+0530 2013	IP: 10.0.0.177 User: john@MYDLPSUPPORT	https://mail.google.com/mail/?ui=	Rule: web rule sales Action: Quarantine Channel: Web	Information Type: PCI-Credit Card	supportfile.pdf
Sun May 5 16:50:58 GMT+0530 2013	IP: 10.0.0.177 User: john@MYDLPSUPPORT	https://mail.google.com/mail/?ui=	Rule: web rule sales Action: Quarantine Channel: Web	Information Type: PCI-Credit Card	hqreport.pdf
Sun May 5 16:50:53 GMT+0530 2013	IP: 10.0.0.177 User: john@MYDLPSUPPORT	https://mail.google.com/mail/?ui=	Rule: web rule sales Action: Quarantine Channel: Web	Information Type: IBAN Account Numbers	IBAN.pdf
Sun May 5 16:50:31 GMT+0530 2013	IP: 10.0.0.177 User: john@MYDLPSUPPORT	https://mail.google.com/mail/?ui=	Rule: web rule sales Action: Quarantine Channel: Web	Information Type: PCI-Credit Card	uenc key_ body
Sun May 5 16:50:16 GMT+0530 2013	IP: 10.0.0.177 User: john@MYDLPSUPPORT	https://mail.google.com/mail/?ui=	Rule: web rule sales Action: Quarantine Channel: Web	Information Type: PCI-Credit Card	uenc key_ body
Sun May 5 16:43:10 GMT+0530 2013	User: david@mydlptest.com	Rcpt To: <mydlpenterprise@yaho <john@mydlptest.com>	Rule: email HR Action: Quarantine Channel: Mail	Information Type: IBAN Account Numbers	data
Sun May 5 16:38:53 GMT+0530 2013	User: david@mydlptest.com	Rcpt To: <mydlprotection@gmail. <bakin@mydlp.com>	Rule: email with bcc Action: Quarantine Channel: Mail	Information Type: keyword confidential	data
Sun May 5 16:36:59 GMT+0530 2013	User: david@mydlptest.com	Rcpt To: <mydlprotection@gmail. <bakin@mydlp.com>	Rule: email with bcc Action: Quarantine Channel: Mail	Information Type: keyword confidential	Confidential -sales document.pdf
Sun May 5 16:12:46 GMT+0530 2013	IP: 10.0.0.177 User: john@MYDLPSUPPORT	c:/Users/john/Desktop/companyn	Rule: discovery rule Action: Archive Channel: Discovery	Information Type: Credit Card Numbers	hqreport.pdf
Sun May 5 16:12:46 GMT+0530 2013	IP: 10.0.0.177 User: john@MYDLPSUPPORT	c:/Users/john/Desktop/companyn	Rule: discovery rule Action: Archive Channel: Discovery	Information Type: Credit Card Numbers	Creditcard-testfile.pdf
Sun May 5 16:09:14 GMT+0530 2013	IP: 10.0.0.177 User: john@MYDLPSUPPORT	(MyDLP)Microsoft XPS Document	Rule: printer rule Action: Quarantine Channel: Printer	Information Type: Credit Card Numbers	hqreport.pdf.xps
Sun May 5 16:08:27 GMT+0530 2013	IP: 10.0.0.177 User: john@MYDLPSUPPORT	(MyDLP)Xerox Global Print Driver	Rule: printer rule Action: Quarantine Channel: Printer	Information Type: Credit Card Numbers	supportfile.pdf.xps

Figure3. DLP Logs.

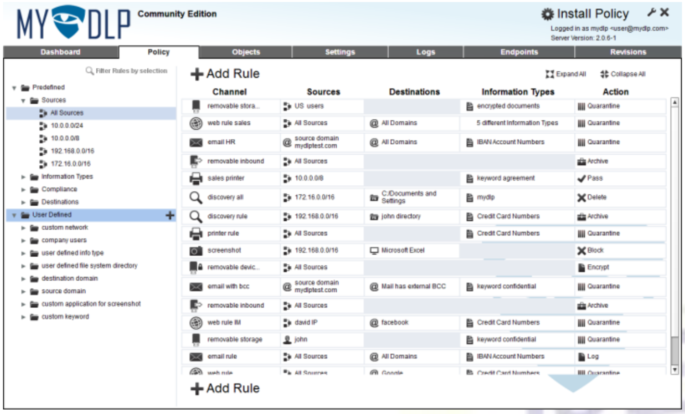
- Admin console basically has different options to perform admin activities.
- Status and overview tab we can find the types of devices and their status (active or Inactive).
- Users and Groups tab gives the functionality to create, delete or/and modify roles / users access to DLP Network, Endpoint and Datacenter tabs, system status is displayed based on device type (Sensor, ICAP server, Grid Worker etc.)
- Notification - automatic alerts are set whenever a device or feature fails to perform the job, i.e. an email alert is sent when any of the devices or services are hit or stop working.
- Settings option gives functionality to set various thresholds support - It opens up a knowledge base for quick help .This helps in saving lot of labor work as in an organization with very huge deployment it is very tidy and uneasy job to keep a track of all the devices and services.
- Content Analysis and Policy Application .All three DLP products make use of content analysis (detection of sensitive content in documents or messages) and application of policy (a specification of how to handle sensitive documents or messages)[9].
- Content Blades:-Content blades are highly accurate pattern-matching detectors of sensitive content. DLP supports two kinds of content blades:

1. Described-content blades are detailed descriptions of sensitive content, and may contain terms, regular expressions, programmatic entities, and other factors to accurately detect classes of sensitive content such as Social Security Numbers. Approximately 150 pre-defined "expert" content blades are available for immediate use in the DLP product, and it can be customized or create other content blades that are unique to organization.
2. Fingerprinted-content blades (or "fingerprints") are mathematical descriptors of individual sensitive documents or fragments of documents. They will "match" any copies of those documents or fragments found anywhere in the organization.

Fingerprints of known sensitive documents are created, and then used to ensure that unauthorized copies of the documents are not being used.

The DLP products use content blades to perform content analysis on intercepted messages, stored files, and files being manipulated by users. Each document or message is assigned a score, or risk factor, depending on how strongly it matches a content blade[10].

Policies



Channel	Sources	Destinations	Information Types	Action
removable stora...	US users		encrypted documents	Quarantine
web rule sales	All Sources	All Domains	5 different Information Types	Quarantine
email HR	source domain mydlptest.com	All Domains	IBAN Account Numbers	Quarantine
removable inbound	All Sources			Archive
sales printer	10.0.0.0/8		keyword agreement	Pass
discovery all	172.16.0.0/16	C:/Documents and Settings	mydlp	Delete
discovery rule	192.168.0.0/16	john directory	Credit Card Numbers	Archive
printer rule	All Sources		Credit Card Numbers	Quarantine
screenshot	192.168.0.0/16	Microsoft Excel		Block
removable devic...	All Sources			Encrypt
email with bcc	source domain mydlptest.com	Mail has external BCC	keyword confidential	Quarantine
removable inbound	All Sources			Archive
web rule IM	david IP	facebook	Credit Card Numbers	Quarantine
removable storage	john		keyword confidential	Quarantine
email rule	All Sources	All Domains	IBAN Account Numbers	Log
web rule	All Sources	Google	Credit Card Numbers	Quarantine

Figure4. DLP Policies.

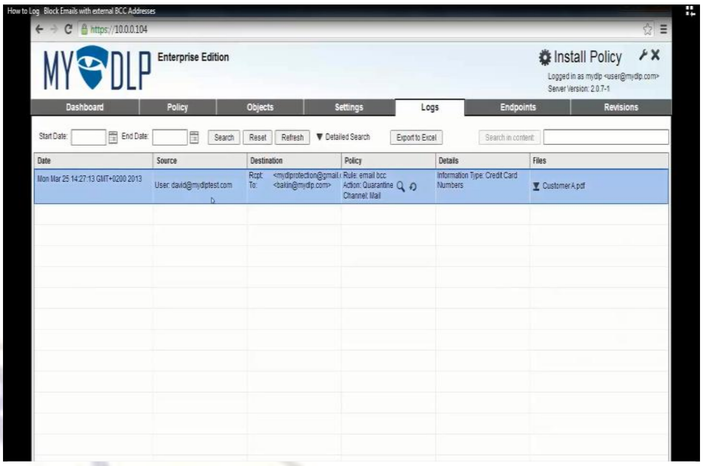
Policies are sets of rules that specify when to create an event (a record that a sensitive document or message has been detected) and how to act on, or remediate, that event. A policy can base its decision on the results of content analysis (the risk factor, or severity, of the analyzed content) and on non-content-based factors such as the identity of the message sender or the destination of the user action. Below figure gives an excerpt of policies in DLP[11].

Approximately 150 pre-defined "expert" policy templates are available for immediate use in the product, however as per the organizational requirement new policies can be created and already existing policies can be customized. Incidents when sufficient number of events occur, DLP creates incidents that a security officer can evaluate and take appropriate steps to manually remediate the security issues that they represent.

There is a dedicated workflow followed to analyze the root cause and follow the remediation process. A dedicated team, working on security incidents handles the workflow. A watch list is maintained and on all the users a vigilant eye is kept, if the security incidents are repeated appropriate action is taken involving other departments like legal, compliance etc.

Below figure shows the high level workflow of Critical Incident Response Center team. The below figure gives an understanding of the types of Incidents / Events, date and time they occurred, severity level, sender or owner details, protocol used, the exact file name or information along with details of type of policy violated.

From the GUI incidents can also be differentiated based on type (Network, Datacenter & Endpoint). As per the requirement it can be filtered with date ranges (day, week, and month)



Date	Source	Destination	Policy	Details	Files
Mon Mar 25 14:27:13 GMT+0200 2013	User: david@mydlptest.com	Rcpt: <mydlpprotection@gmail. To: <bakin@mydlp.com>	Rule: email bcc Action: Quarantine Channel: Mail	Information Type: Credit Card Numbers	Customer A.pdf

Figure5. Blocking mail containing confidential data.

The above figure is an example of how we can prevent data leakage in cloud computing. Here we can observe how confidential mail got blocked when an authorized user try to sent it to any other end points. Because the authorized person is not granted permission to sent confidential data. This is the best example of checking sensitivity of data in cloud computing.

IV. CONCLUSION

In this paper, we do analysis of data leakage prevention. Why it can balance the data security and user convenience. And also suggest the way to prevent it by suing My DLP technology. We‘ve also shown the implementation details of this technology in our experiment part. However, it is very easier to implement DLP technology which will deals will cloud security[12]. Our future work will focus on data leakage analysis in cloud computing using a virtual cloud network.

V. ACKNOWLEDGEMENTS

We grateful to Mrs.Savita Shiwani for her unprecedented guidance, suggestions for this paper.

REFERENCES

[1] *Ma Jun, Wang Zhiying, Ren Jiangchun, Wu Jiangjiang, Cheng Yong and Mei Songzhu," The Application of Chinese Wall Policy in Data Leakage Prevention" in International Conference on Communication Systems and Network Technologies,2012.*

[2] *Charles PEREZ, Babiga BIRREGAH, Marc LEMERCIER, "The Multi-layer imbrication for data leakage prevention from mobile devices" in IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications,2012.*

[3] *Zhang Xiaosong , Liu Fei , Chen Ting , Li Hua , "Research and Application of the*

Transparent Data Encpryption In Intranet Data Leakage Prevention" in International Conference on Computational Intelligence and Security,2009.

[4] Janusz Marecki, Mudhakar Srivatsa, Pradeep Varakantham, "A Decision Theoretic Approachto Data Leakage Prevention" in IEEE International Conference on Social Computing / IEEE International Conference on Privacy, Security, Risk and Trust.

[5] M. Srivatsa, P. Rohatgi, S. Balfe, and S. Reidt, "Securing information flows: A metadata framework," in Proceedings of 1st IEEE Workshop on Quality of Information for Sensor Networks (QoISN), 2008.

[6] D. Roberts, G. Lock, and D. Verma, "Holistan: A Futuristic Scenario for International Coalition Operations," in In Proceedings of Fourth International Conference on Knowledge Systems for Coalition Operations (KSCO), 2007.

[7] Zhao Yong, Liu Jiqian, Han Zhen,Shen Changxiang, "The Application of Information Leakage Defendable Model in Enterprise Intranet", In: Journal of Computer Research and Development, pp761-767 2007 44(5)

[8] Wang Lei, ZHUANG Yi, Pan Long-ping, "Design and implementation of file watching system based on mandatory accesscontrol", In: Computer Application. Vol.26 No.12 Dec.2006

[9] Lei Zheng, Zhao-feng Ma, Ming Gu, "Techniques of File System Filter Driver-based and Security-enhanced Encryption System", In:Journal of Chinese Computer Systems.Vol28.no.7,July 2007

[10] Shufen Liu, Zhagxiang Zhang, Yaorui Cui, Lin tao Wu;"A New information Leakage Defendable Model " In: Computer-Aided Industrial Design and Conceptual Design, 2008. CAID/CD 2008. 9th International Conference on,pp:109-112, Nov. 2008

[11] Microsoft Corporations: "Using Encrypting File System", published: November ,2005

[12] Ulf T. Mattsson, CTO Protegrity, "A Practical Implementation of Transparent Encryption and Separation of Duties in Enterprise Databases, Protection against External and Internal Attacks on Databases", In: E-Commerce Technology, 2005. CEC 2005. Seventh IEEE International Conference on, pp:559–565,19-22 July 2005