DATA LEAK PREVENTION ON SENSITIVE DATA USING LEVENSHTEIN DISTANCE ALGORITHM

R.Bhavani[1] R.Jayashree[2] S.Sushmitha[3] Dr.T.Kalaichelvi [4]

[4]Dr.T.Kalaichelvi,Professor, Dept. of Computer Science & Engineering, Panimalar Institute Of Technology, Chennai,Tamilnadu, India

---

ABSTRACT---*Statistics from security firms, investigate establishments and government associations demonstrate that the quantity of information whole examples has developed quickly lately. Among different information release cases, human missteps are one of the fundamental drivers of information misfortune. According to a report from Risk Based Security (RBS) the amount of discharged delicate data records has extended altogether in the midst of the latest couple of years, i.e., from 412 million in 2012 to 822 million in 2013. Purposefully orchestrated strikes, unexpected breaks (sending mystery messages to unclassified email records) and human oversights (selecting the wrong advantage) incite to most by far of the data spill scenes. Recognizing and preventing data spills requires a game plan of complementary courses of action, which may consolidate data spill disclosure data containment stealthy malware area and approach approval. Organize information spill location (DLD) typically performs profound bundle examination (DPI) and outputs for any occasions of fragile data outlines. In this paper, a data spill area course of action which can be outsourced from affiliation, layout and execute Lucene web look instrument structure Levenshtein-expel strategy to avoid data spill and moreover give security sparing to delicate data.*

Keywords: Risk based security, data leak detection, deep packet inspection, Levenshtein-distance, DLD, DPI, data spill

1. INTRODUCTION

Organization in every industry meets with loss of sensitive data. These data is transferred intentionally or an intentionally by the employees within the organization. Example, leakage of government, financial data, baking records, and many more. Such loss of internal data damages industry standards, brand and reputations. As these data are more valuable one it is necessary to protect from loss in public and data loss detection tools should be designed to avoid such data leakage and thus preventing the loss of confidential data from being leaked out of an industry.For example an organization poses details about their employees. It is essential to protect the data from leakage.In case when those sensitive data about the employee is intentionally transferred out by an employee; it can put the organization at risk when such sensitive information is disclosed to the public. So in practice, industries are struggling to provide right access to the information to the right people in organization.

Data spillage is described as the impromptu or unplanned dispersal of private or delicate data to an unapproved component. Sensitive data in associations and affiliations consolidate intellectual property (IP), cash related information, understanding information, singular charge card data, and other information depending upon the business and the business. Data spillage speaks to a noteworthy issue for associations as the amount of scenes and the cost to those experiencing them continue expanding. Data spillage is enhanced by the way that transmitted data (both inbound and out-bound), including messages, messaging, site structures, and record trades among others, are for the most part unregulated and unmonitored on their way to their objectives. The potential harm and unfriendly results of an information spillage occurrence can be characterized into two classes: Direct and Indirect Losses. Coordinate misfortunes allude to unmistakable harm that is anything but difficult to gauge or to appraise quantitatively. Aberrant misfortunes, then again, are considerably harder to measure and have a substantially more extensive effect regarding cost, place, and time. Coordinate misfortunes incorporate infringement of controls, (for example, those securing client protection) bringing about fines, settlements or

client pay expenses case including claims loss of future deals expenses of examination and healing or rebuilding charges. Aberrant misfortunes incorporate diminished share cost therefore of negative attention harm to an organization's goodwill and notoriety client relinquishment and presentation of protected innovation (marketable strategies, code, money related reports, and meeting plans) to contenders.

The goal is to detect the leak of sensitive data. When a specific data is considered as sensitive then proper data protection control have to be taken by applying effective encryption. In addition to this, policy have to clearly state whether the employee is allowed to send sensitive information to the third parties. Encryption is a powerful tool to secure the information. It protects the information whenever a data is in transit or at rest.

2. RELATED WORK

The blame discovery approach introduced is identified with the information provenance issue [1] following the heredity of S items suggests basically the identification of the liable specialists. Instructional exercise [2] gives a decent outline on the exploration directed in this field. Proposed arrangements are space particular, for example, genealogy following for information stockrooms [3], and accept some earlier learning in transit an information view is made out of information sources. The issue plan with items and sets is more broad and rearranges heredity following, As far as the information portion methodologies are concerned, work is for the most part significant to watermarking that is utilized as a method for setting up unique responsibility for articles. [4], and sound information [5] whose advanced portrayal incorporates extensive repetition. As of late, [6], [7], [8], and different works have likewise considered imprints addition to social information.

With the quick development of database business on the net, the information might be hazardous subsequent to going through the unsecure organize. The information buyers may dither to purchase the information benefit for the accompanying doubt. In the first place, the information beneficiary may speculate that the information are messed with by unapproved individual. Second, they may speculate the information got are not delivered and given by the approved providers. Third, the providers and buyers really with various intrigue ought to have diverse parts of rights in the database administration or utilizing. So how to secure and confirm the information turns out to be essential here. The current surge in the development of the web brings about offering of an extensive variety of electronic administrations, for example, database as an administration, computerized archives and libraries, web based business, online choice emotionally supportive network and so on.

Throughout working together, once in a while touchy information must be given over to as far as anyone knows trusted outsiders. For instance, a healing centred may give persistent records to specialists who will devise new medications. The proprietor of the information the merchant and the as far as anyone knows confided in outsiders the specialists. There is a recognize when the wholesaler's touchy information have been spilled by operators, and if conceivable to distinguish the specialist that released the information. Here consideration of applications where the main sensitive data can't be irritated. Aggravation is a particularly significant framework where the data are balanced and made "less delicate" before being given to administrators. For example, one can add subjective commotion to particular attributes, or one can supplant rectify values by degrees [14]. Regardless, every so often, it is indispensable not to change the primary dealer's data. For example, if an outsourcer is doing our fund, he ought to have the right pay and customer money related adjust numbers. If helpful researchers will treat patients (instead of simply preparing bits of knowledge), they may require exact data for the patients. By and large, spillage disclosure is dealt with by watermarking, e.g., a fascinating code is embedded in each passed on copy. If that copy is later found in the hands of an unapproved party, the leaker can be perceived. Watermarks can be extraordinarily important now and again, yet again, incorporate some modification of the main data. Additionally,

watermarks can generally be demolished if the data recipient is malevolent.

Shu et.al, [9] displayed a security safeguarding data-leak detection (DLD) answer for settle the issue where an uncommon arrangement of delicate information condensations is utilized as a part of location. The benefit of the technique is that it empowers the information proprietor to securely appoint the discovery operation to a semi legit supplier without uncovering the delicate information to the supplier. To depict how Internet specialist co-ops can offer their clients DLD as an extra administration with solid protection ensures. The assessment comes about demonstrate that the strategy can bolster precise identification with modest number of false alerts under different information spill situations.

Outskirts et.al, [10] proposed an approach for measuring data spill limit in system movement. Rather than attempting to identify the nearness of delicate information—an incomprehensible errand in the general case—will likely gauge and compel its greatest volume. Main exploit the knowledge that most system movement is rehashed or controlled by outer data, for example, convention details or messages sent by a server. By sifting this information, a disengage and evaluate genuine data spilling out of a PC. In this paper, there is a exhibit estimation calculations for the Hypertext Transfer Protocol (HTTP), the principle convention for web perusing. At the point when connected to genuine web perusing movement, the calculations could rebate 98.5% of measured bytes and successfully segregate data spills.

3. DATA LEAK PREVENTION

Two most vital players in the proposed model is Data proprietor and Mail server.

(i) Data Owner possesses the touchy information and approves the DLD supplier to investigate the system movement from the hierarchical systems for abnormalities, in particular accidental information spill.

(ii) Mail Server - DLD supplier assesses the system movement for potential information spills. In this paper main aim is to concentrate on recognizing incidental information holes, and they accept the substance in record framework or system movement (over directed system channels) is accessible to the investigation framework. A regulated system channel could be a decoded channel or an encoded divert where the substance in it can be separated and checked by an expert. Authority has the threshold for every categorized position of users

In the security model, a assumption that the analysis system is secure and trustworthy. Privacy-preserved data-leakdetection can be achieved by leveraging special protocols and computation steps. It is another functionality of a detection system,

In this paper the implementation of the web service to maintain the users and sensible content instead of data bases because of static implementation and rough data handling. Even the sensible data storage have to preserved from threatens in existing system.

The proposed system consists of the following parts: (a) Content Outsourcing without DLD (b) Build data leakage detection framework (c) Content Outsourcing with DLD Checker (d) Sensitive Data Detection.

3.1. Phase1: Content Outsourcing without DLD

In this module user's register in mail server with their name, authorized job position and their authorized e-mail domain. And the users can transfer their file using without any restriction of sensible content checking.

There is no content checking and domain filtering on their transformed sensible data. Sensible content is outsourcing from one organization to another organization performed by user. The content can be of any file (text, document). Outsourcing will not reach DLD and directly reach its destination or organization. Here outsourcing mechanism of transferred data is offending over the protocol.

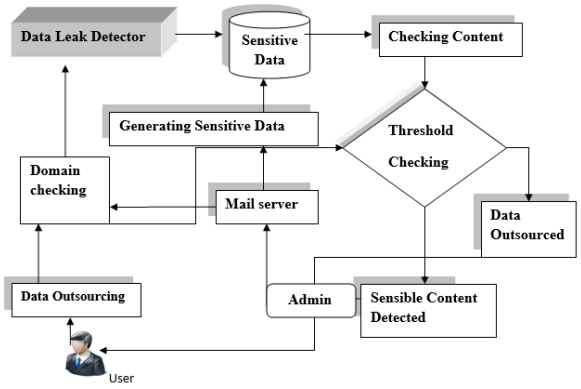


Figure 3.1 Proposed System Architecture

This figure explains how the sensitive data are transferred based on the threshold value.

3.2. Phase2: Build data leakage detection framework:

In this module mail server data owner generates a sensitive data and stored in the cloud and create the directory for lucene search framework and other data leakage detectors. Data owner's cloud contains much sensitive information about their authorized customer's details, information technology source, and database and server details. This sensitive information is maintained by Data Leak Detector. Using this DLD referenced directory perform data leak detection mechanism.The DLD consist of lucene search engine framework, levenshtein distance algorithm and own shuffled checking algorithm. The DLD directly configured with cloud and can refer every data transformation outsourcing from authorized user transformation.

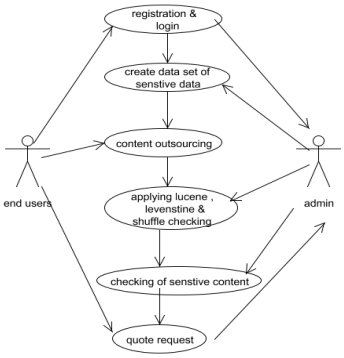


Figure 3.2 Use case diagram of data leak on sensitive data

3.3. Phase3: Content Outsourcing with DLD Checker

DLD is the one will check all the outsourcing content before it transmit to the other organization. All the outsourced contents are check with sensitive data. All the sensitive data are maintaining in index file. Using this index file DLD identify the sensitive data concurrently with domain filtering and threshold assigning based on their email domain. DLD will check every line of the sending data with the sensitive file. DLD will not allow any sensitive data will leak to any of the other organization.

In proxy mail server the every occurrence of transformed contents are filter by users email domain. All users' details are retrieved from the cloud using their email. Then threshold assigned for the users based on their authorized job position and the transferred content has been tested by lucene frameworksearch engine, levenshtein distance checking and shuffling algorithm.

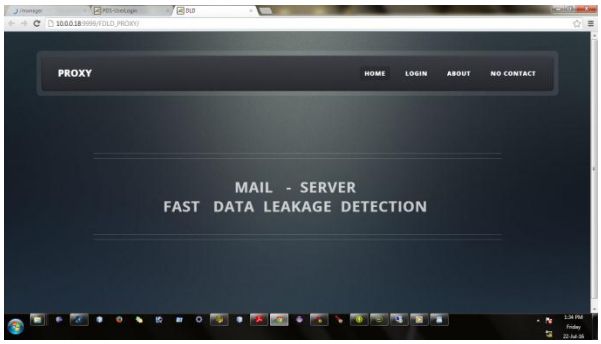
3.4. Phase4: Sensitive Data Detection

Once the DLD framework checks the outsourced content, if any data leak is identified means DLD will detect the sensitive data. Here DLD will check not only the sensitive data and also it will check some access condition. Every data owner maintain common access condition every file. For example, all the contents are encrypted before they outsourced. If DLD identified any sensitive information outsourcing means they will detect the sensible content in between of the file outsourcing.For the purpose of false alert, they maintain threshold of every domain and users position. If the sensible content percentage of transferred file exceeds the threshold percentage which trigger alert mail to Admin of the proxy mail server. Alert mail consists of entire details about the users even what are the sensible contents are pings from the transferred content by the DLD framework.

4. EXPERIMENTAL RESULTS

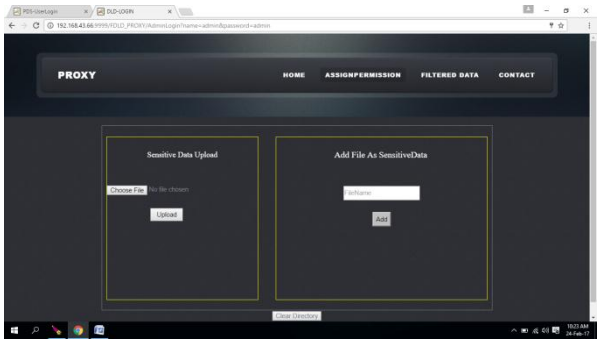
In section IV the experimental results are followed.

4.1. USER LOGIN:



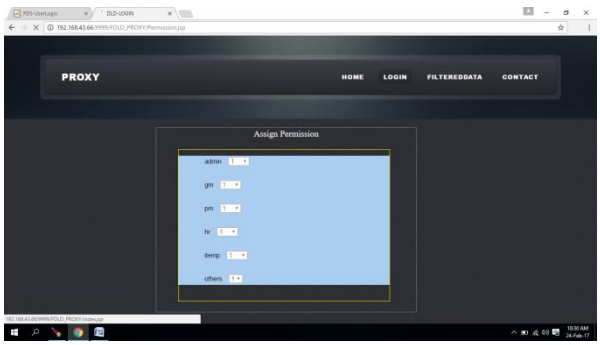
The user use this login to log on to the server and login themselves by giving there details.

4.2.ADDING SENSITIVE DATA:



In this module the admin add the sensitive data to the lucene search engine frame work.

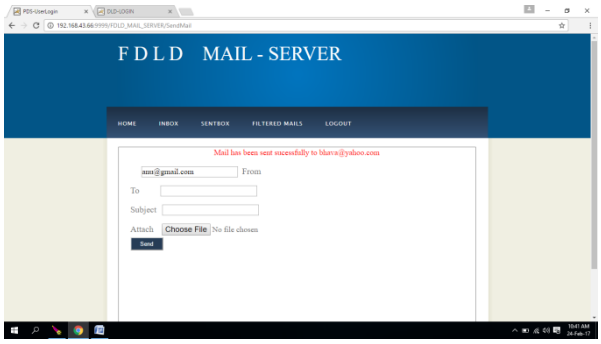
4.3. ASSIGNING THRESHOLD VALUE



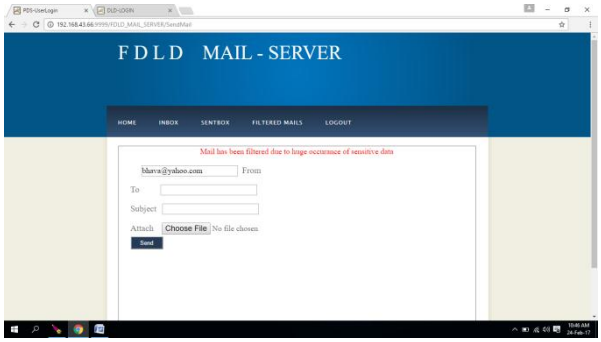
Here in proposed system the threshold value for each employee in the organization is assigned.

4.4. SENDING THE CONTENT OF SENSITIVE DATA:

4.4.1) If the threshold value matched the data is sent



4.4.2) If the threshold value is not matched the data is not sent and transferred to filter mail:



5. CONCLUSION

In this paper, substance review system was introduced for identifying holes of delicate data in the substance of records or system activity. The location approach depends on adjusting two tested successions for similitude examination. The test comes about recommend that the arrangement technique is valuable for distinguishing different basic information spill situations. The parallel renditions of the model give significant speedup and show high adaptability of the outline.

REFERENCES

[1] X. Shu, J. Zhang, D. Yao, and W.-C. Feng, "Rapid and parallel content screening for detecting transformed data exposure," in *Proc. 3rdInt. Work shop Secur. Privacy Big Data (BigSecurity)*, Apr./May 2015,pp. 191–196.

[2] X. Shu, D. Yao, and E. Bertino, "Privacy-preserving detection ofsensitive data exposure," *IEEE Trans. Inf. Forensics Security*, vol. 10,no. 5, pp. 1092–1103, May 2015.

[3]X. Shu, J. Zhang, D. Yao, and W.-C. Feng, "Rapid screening oftransformed data leaks with efficient algorithms and parallel computing,"in*Proc. 5th ACM Conf. Data Appl. Secur. Privacy (CODASPY)*,San Antonio, TX, USA, Mar. 2015, pp. 147–149.

[4] Y. Jang, S. P. Chung, B. D. Payne, and W. Lee, "Gyrus: A framework for user-intent monitoring of text-based networked applications," in *Proc.23rd USENIX Secur. Symp.*, 2014, pp. 79–93.[5] A. Nadkarni and W. Enck, "Preventing accidental data disclosure in
modern operating systems," in *Proc. 20th ACM Conf. Comput. Commun.Secur.*, 2013, pp. 1029–1042.

[6] K. Lee, H. Lin, and W.-C. Feng, "Performance characterization of data-intensive kernels on AMD fusion architectures," *Comput.Sci.-Res. Develop.*, vol. 28, no. 2, pp. 175–184, May 2013.

[7] X. Shu and D. Yao, "Data leak detection as a service," in *Proc. 8thInt. Conf. Secur. Privacy Commun.Netw. (SecureComm)*, Padua, Italy,Sep. 2012, pp. 222–240.

[8] M. Blanton, M. J. Atallah, K. B. Frikken, and Q. Malluhi, "Secure and efficient outsourcing of sequence comparisons," in *Proc. 17th Eur. Symp.Res. Comput. Secur.*, 2012, pp. 505–522.

[9] J. Jang, D. Brumley, and S. Venkataraman, "BitShred: Feature hashing
malware for scalable triage and semantic analysis," in *Proc. 18th ACMConf. Comput. Commun.Secur.(CCS)*, 2011, pp. 309–320.

[10] J. Croft and M. Caesar, "Towards practical avoidance of information
leakage in enterprise networks," in *Proc. 6th USENIX Conf. Hot Topics Secur. (HotSec)*, 2011, p. 7.

[11]A. Squicciarini, S. Sundareswaran, and D. Lin, "Preventing information
leakage from indexing in the cloud," in *Proc. 3rd IEEE Int. Conf. Cloud Comput.*, Jul. 2010, pp. 188–195.

[12] D. Ficara, G. Antichi, A. Di Pietro, S. Giordano, G. Procissi, and F. Vitucci, "Sampling techniques to accelerate pattern matching in network intrusion detection systems," in *Proc. IEEE Int. Conf. Commun.*,May 2010, pp. 1–5.

[13]K. Borders and A. Prakash, "Quantifying information leaks in outbound
In this paperb traffic," in *Proc. 30th IEEE Symp.Secur. Privacy (SP)*, May 2009,
pp. 129–140.

[14] S. Kumar, B. Chandrasekaran, J. Turner, and G. Varghese, "Curing
regular expressions matching algorithms from insomnia, amnesia, and acalculia," in *Proc. 3rd ACM/IEEE Symp. Archit. Netw. Commun.Syst. (ANCS)*, 2007, pp. 155–164.

[15] W. Liu, B. Schmidt, G. Voss, A. Schroder, and W. Muller-Wittig, "Bio sequence database scanning on a GPU," in *Proc. 20th Int. ParallelDistrib. Process.Symp.*, Apr. 2006, pp. 1–8.