RESEARCH ARTICLE OPEN ACCESS

A Survey On Data Leakage Detection

Mrs. Grinal Tuscano*, Ms. Humairah Kotadiya**, Mr. Vikrant Bhat**, Mr. Rollan Fernandes**, Mr. Amod Panchal**
*,**(Department of Information Technology, Mumbai University, St. Francis Institute of Technology, Mumbai)

ABSTRACT
Data is an important assets for an enterprise. Data must be protected against loss and destruction. In IT field huge data is being exchanged among multiple people at every moment. During sharing of the data, there are huge chances of data vulnerability, leakage or alteration. So, to prevent these problems, a survey on data leakage detection system has been done. This paper talks about the concept, causes and techniques to detect the data leakage. Businesses processes facts and figures to turn raw data into useful information. This information is used by businesses to generate and improve revenue at every mile stone. Thus, along with data availability and accessibility data security is also very important.

Keywords – Data Leakage Detection, Water Marking, Guilty Agent, Data Allocation

I. INTRODUCTION

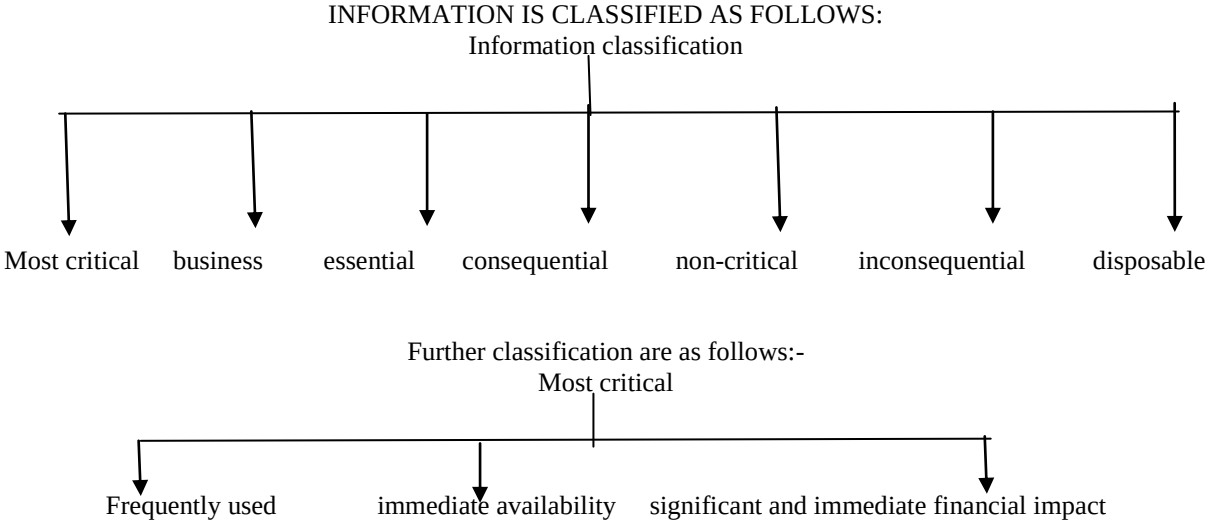
The value of the data is incredible, so it should not be leaked or altered. Though there are number of systems designed for the data security by using different encryption algorithms, there is a big issue of the integrity of the users of those systems. It is very hard for any system administrator to trace out the data leaker among the system users. It creates a lot many ethical issues in the working environment of the office. Therefore Data leakage is the big challenge of the enterprises.

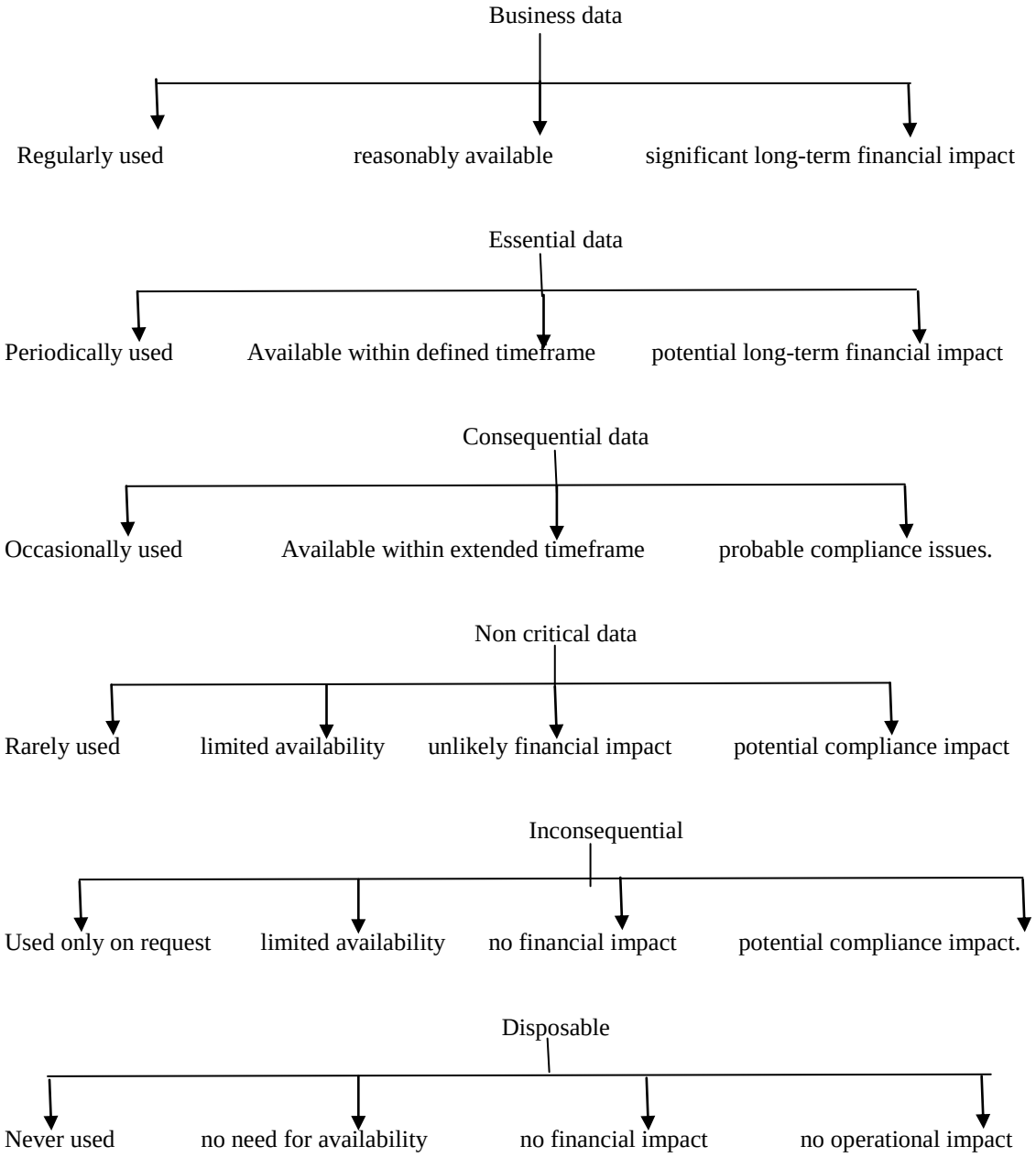
The paper is organized as follows. Data identification, Data classification, Organization of Prevention Techniques, Improved security (measuring outcome). Literature review introduces several techniques used in existing systems along with their shortcomings. Further section shows the completeness

II. LITERATURE REVIEW

Data security is critical for most businesses and even home computer users. Client information, payment information, personal files, bank account details - all of this information can be hard to replace and potentially dangerous if it falls into the wrong hands. Data lost due to disasters such as a flood or fire is crushing, but losing it to hackers or a malware infection can have much greater consequences.

However, some of this information isn't intended to leave the system. The unauthorized access of this data could lead to numerous problems for the larger corporation or even the personal home user. Having your bank account details stolen is just as damaging as the system administrator who was just robbed for the client information in their database.





- Business information: As a business firm it holds large amount of important data which comprises of business strategies, list of customers and their personal details, personal information of the employees, financial records and budget, salary accounting of the employees etc. The above mentioned every information holds a value to the firm and that is the reason the company would ensure the security of such sensitive data. customers and employee details is a kind which can afford alteration so watermarking and fake object techniques can be suitable to prevent them from leakage but on the other hand financial records can be not altered so in this case we can use steganography and data allocation techniques.
- Chemical formulas: Most of the products have like soap, medicines, cosmetics etc have their individual chemical formula which is used for production. Every particular brand wants their products to be unique, suppose consider a medicine brand produces a medicine or more, the producers will never ever want their competitors to defeat them in market for which the chemical formula needs to be kept secure from leakage while distribution. Most importantly these formulas are needed to be accurate therefore it is not affordable to use such techniques which alters the data and there techniques such as data

allocation steganography can be used to prevent data from leakage

- Criminal information: The police department comprises of criminal record or information wherein they have list of top criminal as well less threatening criminals and the records of the crime they have committed. Such records are very vulnerable to attacks like leakage, besides the police department also have strategies to encounter these criminals and this is the reason why they need to secure the information. These records should not be altered in any way and this is the reason steganography, cryptography and data allocation can be used. Since we are talking about crime other name which comes into our minds is digital forensics. Most of the people have an opinion that digital forensics means investigation of crime. There are three basic and essential principles in digital forensics: that the evidence is acquired without altering it; that this is demonstrably so; and that analysis is conducted in an accountable and repeatable way. The definition of digital forensics says 'The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions.

- Military information: offensive strategies, defensive strategies, weaponry and details of any particular soldier are mainly in the military information. These information cannot be affordable to be found by any non military individual or any enemy since it can lead to defeat of that particular military force who data has been leaked and this is the scenario where cryptography comes into picture to secure the information

PREVENTION TECHNIQUES:

Let us discuss about the prevention and then later come to the briefing of techniques

A. Encryption

Encryption has become a critical security feature for thriving networks and active home users alike. This security mechanism uses mathematical schemes and algorithms to scramble data into unreadable text. It can only by decoded or decrypted by the party that possesses the associated key.

(FDE) Full-disk encryption offers some of the best protection available. This technology enables you to encrypt every piece of data on a disk or hard disk drive. Full disk encryption is even more powerful when hardware solutions are used in conjunction with software components. This combination is often referred to as end-based or end-point full disk encryption.

B. **Strong User Authentication**

Authentication is another part of data security that we encounter with everyday computer usage. Just think about when you log into your email or blog account. That single sign-on process is a form authentication that allows you to log into applications, files, folders and even an entire computer system. Once logged in, you have various given privileges until logging out. Some systems will cancel a session if your machine has been idle for a certain amount of time, requiring that you prove authentication once again to re-enter.

The single sign-on scheme is also implemented into strong user authentication systems. However, it requires individuals to login using multiple factors of authentication. This may include a password, a one-time password, a smart card or even a fingerprint.

Personal data : This critical information is powerful ammunition to an identity thief as it could give them easy access to your bank account, credit cards, medical records and other critical assets.An organisation holds highly sensitive or confidential personal data (such as information about individuals' health or finances) which could cause damage or distress to those individuals if it fell into the hands of others. The organisation's information security measures should focus on any potential threat to the information or to the organisation's information systems.

As part of its security measures, an organisation ensures that information on laptop computers issued to staff is protected by encryption, and that desk-top computer screens in its offices are positioned so that they cannot be viewed by casual passers-by. Paper waste is collected in secure bins and is shredded on site at the end of each week.

This paper focuses on prevention of sensitive data as well as detection of any data been leaked. There are two basic characterization of techniques first where we need to alter the data and second without alteration

1. Techniques with alteration
   - Watermarking
   - Fake objects
   - Cryptography
2. Techniques without alteration
   - Steganography
   - Data allocation

Let us now study about these techniques in brief.

1. Water marking technique

Traditionally, leakage detection is handled by watermarking, e.g., a unique code is embedded in each distributed copy. If that copy is later discovered

in the hands of an unauthorized party, the leaker can be identified. Watermarks can be very useful in some cases. If we consider Real time example then watermarking can also be used for compressed video data. To provide copy protection and copyright protection for digital audio and video data, two complementary techniques are being developed: encryption and watermarking. Encryption techniques can be used to protect digital data during the transmission from the sender to the receiver. However, after the receiver has received and decrypted the data, the data is in the clear and no longer protected. Watermarking techniques can complement encryption by embedding a secret imperceptible signal, a watermark, directly into the clear data. This watermark signal is embedded in such a way that it cannot be removed without affecting the quality of the audio or video data. The watermark signal can for instance be used for copyright protection as it can hide information about the author in the data. The watermark can now be used to prove ownership in court. Another interesting application for which the watermark signal can be used is to trace the source of illegal copies by means of fingerprinting techniques.

Disadvantage-Watermarks involve some modification of the original data. Furthermore, watermarks can sometimes be destroyed if the data recipient is malicious.

2. Fake object technique

Fake object are basically alteration in original data which apparently improves the probability To find guilty agents. The distributor may be able to add fake objects to the distributed data in order to improve his effective-ness in detecting guilty agents. The idea of disturbing or modifying data to detect leakage is not new. However, in most cases, individual objects are modified, e.g., by adding random noise to sensitive salaries, or adding a watermark to an image. Speaking about real time use of fake objects our study says we can consider trace records one of them. Trace records are basically owned addresses by entities. Let us consider two companies X & Y, now suppose company X sells a mailing list to company Y for say advertisement. At that time company X adds trace records to the mailing list because of which everytime when company Y uses that mailing list company X receives a copy of it. This therefore can identify unauthorized use of data.

Disadvantage- it leads to alteration of original data. Altering original data may not be suitable everytime for e.g suppose we have Financial information like budget or employee's salary, such sensitive information cannot be altered as any alteration can lead to financial crises of the company.

3. Steganography:

The art and science of hiding information by embedding messages within other, seemingly harmless messages. Steganography sometimes is used when encryption is not permitted. Or, more commonly, steganography is used to supplement encryption. An encrypted file may still hide information using steganography, so even if the encrypted file is deciphered, the hidden message is not seen. Steganography can be considered as technique which embeds sensitive information with other data using rules and techniques because of which we can identify authorized person or entity. Apparently watermarking mostly widely used in steganography. Speaking in brief about watermarking, it has two types visible and invisible watermarking. Visible watermarking as the name suggest is visible it is a text or logo which depicts the ownerships. For example the logo of famous company like BMW, Audi, Mercedes very apparently indicates their particular cars. These logo can even be for advertisement and promotion of these firm. Coming to invisible watermarks, they can be embedded in audio, video, images as well as text. The remarkable aspect about invisible marking is that they appears to be the original object and besides this technique can be used for copyright prevention which authorizes authors, creators, writers etc. There is list of Real time application of steganography like in the business world it can used to hide secret chemical or plans for new inventions. Even terrorist can use steganography for secret communication and attack plans. Traditionally steganography was in used while making maps where the cartographers adds tiny fictional street to prevent the map from copycats. Disadvantage: Message is hard to recover if the image is subjected to attack such translation and rotation besides it relative easy to detect

4. Cryptography

Network security is one the major issues and data security today strongly requires encryption techniques to maintain the integrity, confidentiality of the information. The information can be hacked by any intruder and can be used for many malicious purposes. To secure the data many cryptography techniques are available such as diffie hellman, AES, Hash etc. cryptography is usually conducted between two different entities in which one entity sends an encrypted information and the other entity converts the encrypted message to its original form using a secret or private key. The major advantage of this technique is any intruder trying to leak the sensitive data may not understand the encrypted text or the encryption format easily. Ultra-Secure Voting With political upheaval and accusations of voter fraud rampant in developed and developing countries alike, it's clear that making the voting process more secure

is a necessity. Since 2007, Switzerland has been using cryptography to conduct secure online voting in federal and regional elections. In Geneva, votes are encrypted at a central vote-counting station. Then the results are transmitted over a dedicated optical fiber line to a remote data storage facility. The voting results are secured via quantum cryptography, and the most vulnerable part of the data transaction (when the vote moves from counting station to central repository) is uninterruptible. This technology will soon spread worldwide, as many other countries face the specter of fraudulent elections. Disadvantage: It takes long to create the code. If you were to send a code to another person the encryption and the decryption technique usually takes a long time. Overall cryptography is a very long process.

5. Data allocation

The main focus of the data allocation problem as how can the distributor intelligently give data to agents in order to improve the chances of detecting a guilty agent, Admin can send the files to the authenticated user, users can edit their account details etc. Agent views the secret key details through mail. In order to increase the chances of detecting agents that leak data. Coming to the brief description of our proposed system, suppose the admin sends data or file to user at that particular time system generates 4 digit random number also known as secret key and mails it to intended user, the user needs to enter the key while downloading the data or file this apparently ensure the security of the sensitive data. Now if someone other than the intended user tries to acquire the data and enters wrong secret key the system acknowledges the intended user via mail. The intended user now can lock the sensitive file by setting desired password.

COMPARISON

PARAMETERS	WATER MARKING TECHNIQUE	FAKE OBJECT	STEGANOGRAPHY	CRYPTOGRAPHY	DATA ALLOCATION
accuracy	low	low	moderate	low	high
detection	Not easy to detect	Not easy	Not easy to det ect because to find steganographic image is hard.	Not easy to detect ,depend on technology used to generate	
complexity	moderate	moderate	moderate	high	low
robust	yes	no	yes	yes	yes
strength	Extend information and become an attribute of the cover image	Enhances the probability to detect the leakage by including the fake data	Hide message without altering message, it conceals information	Hide message by altering the message by assigning encryption key	No modification of data
capacity	Capacity depends on the size of hidden data	Depending on the hidden data	Differs as different Technology usually low hiding capacity	Capacity is so high, but as message is long it chances to be decrypt	Low capacity

III. CONCLUSION

From this Study we conclude that the data leakage detection system model is very useful as compare to the existing watermarking model. We can provide security to our data during its distribution or transmission and even we can detect if that gets leaked. Thus, using this model security as well as tracking system is developed. Watermarking can just provide security using various algorithms through encryption, whereas this model provides security plus detection technique. This model is very helpful in various industries, where data is distribute through

any public or private channel and shred with third party. Now, industry & various offices can rely on this security & detection model.

Acknowledgements

It is with a great sense of gratitude that I acknowledge the support, time to time suggestions and highly indebted to my guide Ms.Grinal Tuscano

REFERENCES

Journal Papers:

[1] A novel data leakage detection,Priyanka Barge, Pratibha Dhawale, Namrata Kolashetti3 Ass. Prof., Department of Computer Engineering, NIRMALA CHOUHAN International Journal of Modern Engineering Research (IJMER) Vol.3, Issue.1, Jan-Feb. 2013 pp-538-540 ISSN: 2249-6645

[2] Data leakage detection,Sandip A. Kale, Prof. S.V.Kulkarni Department Of CSE, MIT College of Engg, Aurangabad, Dr.B.A.M.University, Aurangabad (M.S), India International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 9, November 2012

[3] Data Leakage Detection, Panagiotis Papadimitriou, Hector Garcia-Molina. January 10,2009

[4] Network security using cryptographic techniques, Sumedha Kaushik, Ankur Singhal, Department of ECE & M.M.unversity Ambala (Haryana) India, volume 2, issue 12, December 2012, ISSN:2277 128X

[5] Information Hiding in Images Using Steganography Techniques *Ramadhan Mstafa, Christian Bach* 2013 ASEE Northeast section conference, Norwich university, reviewed paper, March 14-16, 2013

[6] Digital forencis and Preservation, Jeremy Leighton John, DPC technology watch report 12-03 November 2012.

[7] Steganography and its application security, Ronak Doshi, Pratik Jain, Lalit Gupta, Department of Electonics and Telecommunication, Pune University, India.

[8] Steganography, Cryptography, Watermarking: A Comparative Study Hardikkumar V. Desai (B.Sc., MCA)Research Scholar, Singhania University, Volume 3, No. 12, December 2012 *Journal of Global Research in Computer Science*