Research Paper
On

Data Leakage Detection using Cloud Computing

Submitted By:
Yogendra Singh
15SCSE105097
IBM-Cloud 1

Submitted to:
Mr. Gautam Kumar
Asst. Prof.

Project Title:

Data Leakage Detection Using Cloud Computing

Abstract:

The major aspect of the project is to avoid the data being leaked due to agents or the employees having authorized privilege to access the sensitive data of the company.
Employees within the company play a major role in this as they have physical access to the sensitive data of the company.
Thus the problem converges to a point where there is a strong need to avoid the leakage of data by the agents or the employees.Now-a-days trusted parties will be given sensitive data by the data distributor. So, this data must be protected and should notbe found in an unauthorized place. This paper mainly focuses on a survey of the leakage of the data by the various agents within the organization and the different techniques to avoid this leakage. In several fields of research, data is highly confidential. Since, the data is shared by a large number of people; it is vulnerable to alteration and leakage. So, this paper mainly focuses on a survey of various data leakage detection techniques and the approach proposed mainly focuses on delegated access control with the tracing of MAC address so that the leaker can be detected and the data can be blocked to the unauthorized world.

This paper contains the results of implementation of Data Leakage Detection Model. Currently watermarking
technology is being used for the data protection. But this technology doesn't provide the complete security against date leakage.
This paper includes the difference between the watermarking & data leakage detection model's technology. This paper leads for
the new technique of research for secured data transmission & detection, if it gets leaked.

INTRODUCTION:

OBJECTIVE:

The main objective of this project is to reduce or eliminate the chances of getting the data leaked by anyone (agents here). If data leakage can be prevented and detected using cloud computing then there would be very high security of data and integrity will be safe

Due to leakage of the sensitive data the victim company can collapse to to the exposure of business startergies to the rivalary companies.

Data leakage is the big challenge in front of the industries & different institutes. Though there are numberof systems designed for the data security by usingdifferent 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.

The data leakage detection industry is very heterogeneous as it evolved out of ripe product lines of leading IT security vendors. A broad arsenal of enabling technologies such as firewalls, encryption, access control, identity management, machine learning content based detectors and others have already been incorporated to offer protection against various facets of the data leakage threat. The competitive benefits of developing a "one-stop-shop", silver bullet data leakage detection suite is mainly in facilitating effective orchestration of the aforementioned enabling technologies to provide the highest degree of protection by ensuring an optimal fit of specific data leakage detection technologies with the "threat landscape" they operate in. This landscape is characterized by types of leakage channels, data states, users, and IT platforms.

We consider applications where the original sensitive data cannot be perturbed. Perturbation is avery useful technique where the data is modified and made less sensitive before being handed toagents. For example, one can add random noise to certain attributes, or one can replace exact values byranges. However, in some cases it is important not to alter the original distributor’s data. For example, if an outsourcer is doing our payroll, he must have the exact salary and customer identification numbers. If medical researchers will be treating patients (as opposed to simply computing statistics),they may need accurate data for the patients.
Traditionally, leakage detection is handled by watermarking,a unique code is embedded in eachdistributed copy. If that copy is later discovered in the hands of an unauthorized party, the leaker canbe identified. Watermarks can be very useful in some cases, but

again, involve some modification of theoriginal data. Furthermore, watermarks can sometimes be destroyed if the data recipient is malicious.

In this paper we study unobtrusive techniques for detecting leakage of a set of objects or records.

Specifically, we study the following scenario: After giving a set of objects to agents, the distributordiscovers some of those same objects in an unauthorized place.At this point the distributor canassess the likelihood that the leaked data came from one or more agents, as opposed to having been independently gathered by other means.

Sensitive data of companies and organization includes intellectual property, financial information, patient information, personal credit card data and other information depending upon the business and the industry. A data distributor has given this sensitive data to a set of supposedly trusted agents (third parties). Some of the data are leaked and found in an unauthorized place. The distributor must assess the likelihood that the leaked data came from one or more agents, as opposed to having been independently gathered by other means. We call the owner of the data the distributor and the supposedly trusted third parties the agents. Our goal is to detect when the distributor's sensitive data have been leaked by agents, and if possible to identify the agent that leaked the data. We propose data allocation strategies hat improve the probability of identifying leakages. These methods do not rely on alterations of the released data.In some cases, we can also inject realistic but fake‖ data records to further improve our chances of detecting leakage and identifying the guilty party.

BACKGROUND:

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, but again, involve some modification of the original data. Furthermore, watermarks can sometimes be destroyed if the data recipient is malicious. E.g. A hospital may give patient records to researchers who will devise new treatments. Similarly, a company may have partnerships with other companies that require sharing customer data. Another enterprise may outsource its data processing, so data must be given to various other companies. We call the owner of the data the distributor and the supposedly trusted third parties the agents.

A more important question…how was the data leaked?

The classification by leakage channel is important in order to know how the incidents may be prevented in the future and can be classified as physical or logical.
Physical leakage channel means that physical media containing sensitive information or the document itself was moved outside the organization. This more often means that the control over data was lost even before it leaved the organization.

- In the course of doing business, sometimes data must be handed over to trusted third parties.
- Sometimes these trusted third parties may act as points of data leakage.
- Owner of data is termed as the distributor and the third parties are called as the agents .
- In case of data leakage, the distributor must assess the likelihood that the leaked data came from one or more agents

Current Systems:

Watermark is a signal that is securely, imperceptibly, and robustly embedded into original content such as an image, video, or audio signal, producing a watermarked signal and it describes information that can be used for proof of ownership or tamper proofing.
Privacy concerns exist wherever personally identifiable information is collected and stored in digital form or otherwise. Improper or non-existent disclosure control can be the root cause for privacy issues.

We describe a digital watermarking method for use in audio, image, video and multimedia data. We argue that a watermark must be placed in perceptually significant

components of a signal if it is to be robust to common signal distortions and malicious attack.

DCT-Based Watermarking:

The image is first divided into 8 × 8 pixel blocks. After DCT transform and quantization, the midfrequency range DCT coefficients are selected based on a Gaussian network classifier. The mid-frequency range DCT coefficients are then used for embedding.
Those coefficients are modified using a linear DCT constraints. It is claimed that the algorithm is resistant to JPEG compression.

Spread SpectrumWatermarking:

Spread spectrum is used to embed the watermark in the frequency components of the host image. First the Fourier Transform is applied to the host image is inserted to obtain a modified values .The scaling parameter α is used to determine the embedding strength of the watermark. Different spectral components exhibit different tolerance to modification.
To verify the presence of the watermark, the cross correlation value between the extracted watermark and the original watermark Here, we call the cross correlation the similarity.Experimental results showed that this method resists
JPEG compression with a quality factor down to 5%, scaling, dithering, cropping and collusion attacks.

Wavelet Based Watermarking:

The multi resolution data fusion is used for embedding where the image and the watermark are both transformed into the discrete wavelet domain. The watermark is embedded into each wavelet decomposition level of the host image. During detection, the watermark is an average of the estimates from each resolution level of wavelet decomposition. This algorithm is robust against JPEG compression, additive noise and filtering operations.

Robust Watermarking Technique:

Contrary to the LSB approach, the key to making a watermark robust is that it should be embedded in the perceptually significant components of the image.
A good watermark is one which takes into account the behavior of human visual system. For the spread spectrum based watermarking algorithm, a scaling factor can be used to control the amount of energy a watermark has. The watermark energy should be strong enough to withstand possible attacks and distortions. Meanwhile large watermark energy will affect the visual quality of the watermarked image. A perceptual model is needed toHowever, it is well known that modification of these components can lead to perceptual degradation of the signal. To avoid this, we propose to insert a watermark into the spectral components of the data using techniques analogous to spread spectrum communications, hiding a narrow band signal in a wideband channel that is the data.

The watermark is difficult for an attacker to remove, even when several individuals conspire together with independently watermarked copies of the data. It is also robust to common signal and geometric distortions such as digital-to-analog and analog-to-digital conversion, resampling, quantization, dithering, compression, rotation, translation, cropping and scaling. The same digital watermarking algorithm can be applied to all three media under consideration with only minor modifications, making it especially appropriate for multimedia products. Retrieval of the watermark unambiguously identifies the owner, and the watermark can be constructed to make counterfeiting almost impossible. We present experimental results to support these claims.

Invisible Watermarking:

This technique presents a novel invisible robust watermarking scheme for embedding and extracting a digital watermark in an image. The novelty lies in determining a perceptually important sub image in the host image. Invisible insertion of the watermark is performed in the most significant region of the host image such that tampering of that portion with an intention to remove or destroy will degrade the esthetic quality and value of the image. One feature of the algorithm is that this sub image is used as a region of interest for the watermarking process and eliminates the chance of watermark removal. Another feature of the algorithm is the creation of a compound watermark using the input user watermark (logo) and attributes of the host image. This facilitates the homogeneous fusion of a watermark with the cover image, preserves the quality of the host image, and allows robust insertion-extraction. Watermark creation consists of two distinct phases.
During the first phase, a statistical image is synthesized from a perceptually important sub image of the image. A compound watermark is created by embedding a watermark (logo) into the statistical synthetic image by using a visible watermarking technique. This compound watermark is invisibly embedded into the important block of the host image. The authentication process involves extraction of the perceptive logo as well statistical testing for two-layer evidence. Results of the experimentation using standard benchmarks demonstrates the robustness and efficacy of the proposed watermarking approach.
Ownership proof could be established under various hostile attacks.

unintentional release of secure information to an un trusted environment that is a data distributor has given sensitive data to a set of supposedly trusted agents (third parties) and after giving a set of data objects to agents, the distributor discovers some of those same objects in an unauthorized place and now the goal is to estimate the likelihood that the leaked data came from the agents as opposed to other sources. Not only to estimate the likelihood the agents leaked data, but would also like to find out if one of them in particular was more likely to be the leaker.
Using the data allocation strategies, the distributor intelligently give data to agents in order to improve the chances of detecting guilty agent. Fake objects are added to identify the guilty party. If it turns out an agent was given one or more fake objects that were leaked, then the distributor can be more confident that agent was guilty and when the

distributor sees enough evidence that an agent leaked data then they may stop doing business with him, or may initiate legal proceedings.

Furthermore, watermarks can sometimes be destroyed if the data recipient is malicious. A hospital may give patient records to researchers who will devise new treatments. Similarly, a company may have partnerships with other companies that require sharing customer data. Another enterprise may outsource its data processing, so data must be given to various other companies. We call the owner of the data the distributor and the supposedly trusted third parties the agents.

Drawbacks of Watermarking:

- It involves some modification of data that is making the data less sensitive by altering attributes of the data.

- The second problem is that these watermarks can be sometimes destroyed if the recipient is malicious.

DESIGN OF PROJECT AND ITS IMPLEMENTATION:

Detecting Challenges:

Encryption: and preventing data leaks in transit are hampered due to encryption and the high volume of electronic communications. While encryption provides means to ensure the confidentiality, authenticity and integrity of the data, it also makes it difficult to identify the data leaks occurring over encrypted channels. Encrypted emails and file transfer protocols such as SFTP imply that complementary DLP mechanisms should be employed for greater coverage of leak channels. Employing data leak prevention at the endpoint – outside the encrypted channel – has the potential to detect the leaks before the communication is encrypted.

Access Control: Access control provides the first line of defense in DLP. However, it does not have the proper level of granularity and may be outdated. While access control is suitable for data at rest, it is difficult to implement for data in transit and in use. In other words, once the data is retrieved from the repository, it is difficult to enforce access control. Furthermore, access control systems are not always configured with the least privilege principle in mind. For example, if an access control system grants full access to all code repositories for all programmers, it will not effectively detect data leaks where a programmer accesses a project that he/she is not involved in.

THE CLIPBOARD TECHNIQUE(Proposed System):

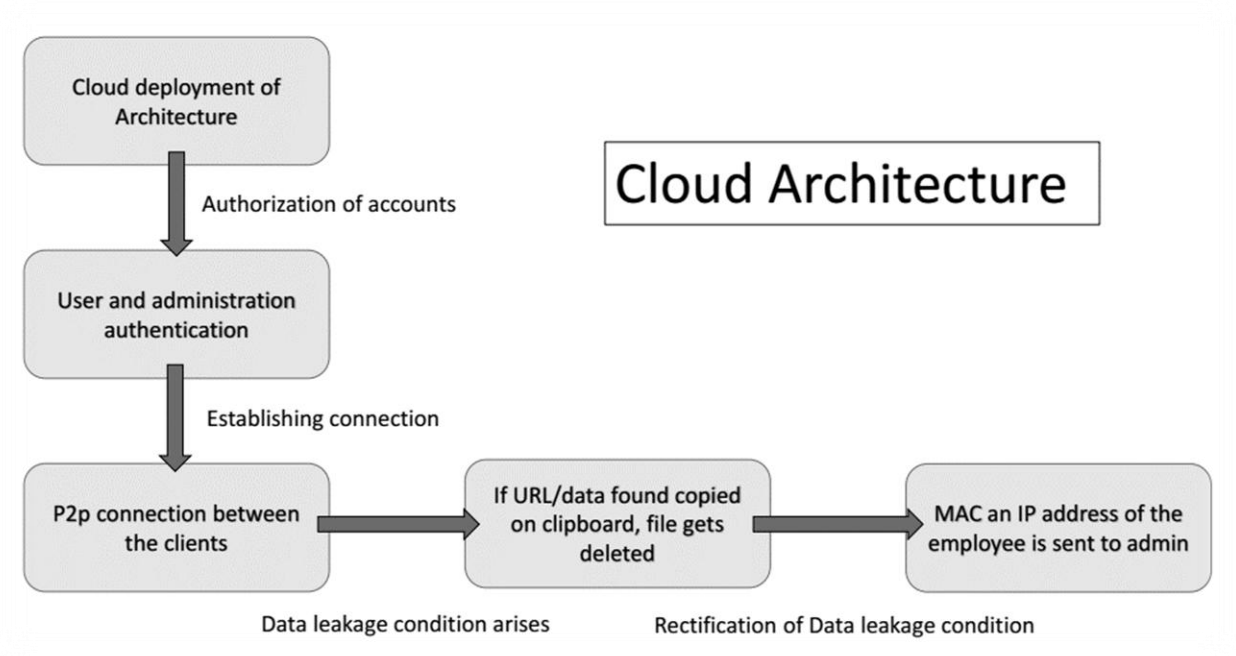
Introduction To Cloud Computing:

CLOUD COMPUTING FEATURES

- Cloud is large group of interconnected computers.
- User centric
- Task centric
- Powerful
- Accessible
- Intelligent
- programmable

Key to the definition of cloud computing is the ―cloud‖ itself. For our purposes, The cloud is a large group of interconnected computers. These computers can be personal computers or network servers; they can be public or private. For example, Google hosts a cloud that consists of both smallish PCs and larger servers. Google's cloud is a private

on(that is, Google owns it) that is publicly accessible (by Google's users). This cloud of computers extends beyond a single company or enterprise. The applications and data served by the cloud are available to broad group of users, cross-enterprise and cross-platform. Access is via the Internet. Any authorized user can access these docs and apps from any computer over any Internet connection.



Data Leakage Detection Architecture

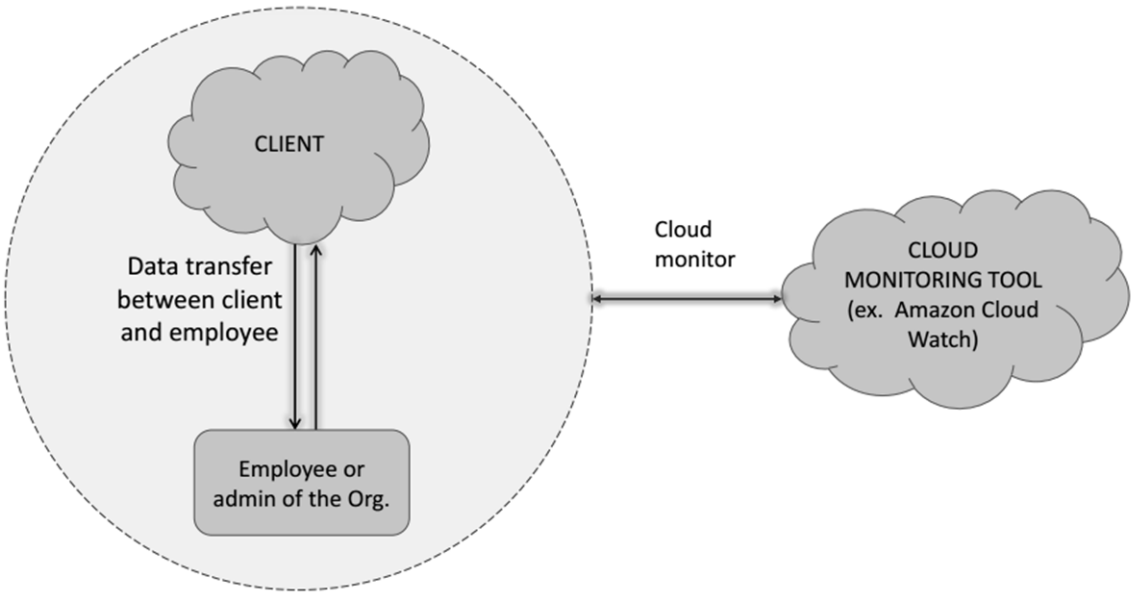
Using Cloud to prevent Data Leakage:

- The data owner first registered into the cloud account. Each and every user has to registered into the cloud. Now the data owner and user will become the authorized person.
- The data owner will upload the file into the cloud. Now the data owner login into the account, at that time the cloud provider verify the already registered owner or not.
- If they are registered owner and then they will transfer the file to the registered user. Now the registered user login into their account the cloud provider again verify the registered user.

- The registered user will download the file sent by the data owner. If someone try to copy the URL, the data get leaked in someone's laptop. Now the details about the unauthorized person will be tracked.
- This tracked information sent as a mobile intimation to the data owner. The mobile intimation will hold informationslike IP address, MAC address and GPS location.

System Requirements(Software/Hardware):

- Intel core i5, core i7 processors which supports intelvt-x technology.
- Minimum 8 GB RAM to deal with huge data efficiently.
- Hard disk drive (~1 TB) for large data storage and exchange.
- A good GPU for high performance/quality graphical interface (NVIDIA will be a good choice).



Cloud Monitoring for Data Leakage

Methodology:

The methods used in this project are:

- Both user and the distributor have to create an account over a cloud service which would establish a secure p2p connection between them.
- The data can be exchanged using the cloud service.If the data is found to be copied or he u, then the data will be removed/deleted using the program created.
- Even if the data passes over to the agent, then the IP address, GPS location and the physical address of the agent will be immediately sent to the CSP (cloud service provider) and the distributor.This way we can save the data from being leaked and the copy of data will be available on the distributor's side.
- The data owner first registered into the cloud account. Each and every user has to registered into the cloud. Now the data owner and user will become the authorized person.
- The data owner will upload the file into the cloud. Now the data owner login into the account, at that time the cloud provider verify the already registered owner or not.

RESULTS :

The registered user will download the file sent by the data owner. If someone try to copy the URL, the data get leaked in someone's laptop. Now the details about the unauthorized person will be tracked

- This tracked information sent as a mobile intimation to the data owner. The mobile intimation will hold informations like IP address, MAC address and GPS location.
- If they are registered owner and then they will transfer the file to the registered user. Now the registered user login into their account the cloud provider again verify the registered user.
- Thus,by sing the above technique we were able to find the guilty employee or the person that was involved in tha data leakage.

CONCLUSIONS:

The Clipboard technique proposed by us is pretty much worth investing as it implements the drawbacks of the previous techniques that were used to avoid data leakage.

Identifying leakage beforehand and correcting for it is an important part of improving the definition of a machine learning problem. Many forms of leakage are subtle and are best detected by trying to extract features and train state-of-the-art models on the problem. This means that there are no guarantees that competitions will launch free of leakage, especially for Research competitions

From this study, I conclude that the data leakage detection industry is very heterogeneous as it evolved out of ripe product lines of leading IT security vendors. A broad arsenal of enabling technologies such as firewalls, encryption, access control, identity management, machine learning content/context-based detectors and others have already been incorporated to offer protection against various facets of the data leakage threat.

The competitive benefits of developing a one-stop-shop, silver bullet data leakage detection suite is mainly in facilitating effective orchestration of the aforementioned enabling technologies to provide the highest degree of protection by ensuring an optimal fit of specific data leakage detection technologies with the "threat landscape" they operate in. This landscape is characterized by types of leakage channels, data states, users, and IT platforms. I also consider, the option of adding “fake” objects to the distributed set. Such objects do not correspond to real entities but appear realistic to the agents. In a sense, the fake objects act as a type of watermark for the entire set, without modifying any individual members. If it turns out that an agent was given one or more fake objects that Ire leaked, then the distributor can be more confident that agent was guilty.

FURTHER WORK:

Furthermore related to the insight of the project ,the project can be further developed into a tool for monitoring the data flow and used for the data mining as well.The tool has many capabilities that can be used to monitor the physical data leakage as well.

REFERENCES:

- [1] SandipA.Kale, Prof. Kulkarni S.V. *(Department Of Computer Sci. &Engg,MIT College of Engg, Dr.B.A.M.University, Aurangabad(M.S), India,* Data Leakage Detection: A Survey, *(IOSR Journal of Computer Engineering (IOSRJCE)ISSN : 2278-0661 Volume 1, Issue 6 (July-Aug 2012), PP 32-35* www.iosrjournals.org
- [4] J.J.K.O.Ruanaidh, W.J.Dowling, and F.M.Boland," Watermarking Digital Images For Copyright Protection", IEE Proc.Vision,Signal and Image Processing,vol.143,no.4,pp.250-256,1996.
- [5] F.Hartung and B.Girod,"Watermarking of Uncompressed and Compressed Video," Signal Processing, vol.66, no.3,pp.283-301,1998.
- [6] S.Czerwinski, R.Fromm,and T.Hodes,"Digital Music Distribution and Audio watermarking," http://www.Scientificcommons.org/43025658,2007.