International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)
International Conference on Information Systems and Computing (ICISC-2013), INDIA.

DATA LEAKAGE DETECTION USING CLOUD COMPUTING

V. Shobana[1], M. Shanmugasundaram[2]
[1]M.E Student, [2]Assistant Professor Department Of CSE ,Velammal Engineering College / AnnaUniversity, Chennai,India;
Email:vshobana88@gmail.com, mshans@gmail.com

Abstract

In the virtual and widely distributed network, the process of handover sensitive data from the distributor to the trusted third parties always occurs regularly in this modern world. It needs to safeguard the security and durability of service based on the demand of users. The idea of modifying the data itself to detect the leakage is not a new approach. Generally, the sensitive data are leaked by the agents, and the specific agent is responsible for the leaked data should always be detected at an early stage. Thus, the detection of data from the distributor to agents is mandatory. This project presents a data leakage detection system using various allocation strategies and which assess the likelihood that the leaked data came from one or more agents. For secure transactions, allowing only authorized users to access sensitive data through access control policies shall prevent data leakage by sharing information only with trusted parties and also the data should be detected from leaking by means of adding fake record`s in the data set and which improves probability of identifying leakages in the system. Then, finally it is decided to implement this mechanism on a cloud server.

***Keywords*-- data leakage, data security, fake records, cloud environment**.

I. INTRODUCTION

The company's Information security depends on employees by learning the rules through training and awareness-building sessions. However, security must go beyond employee knowledge and cover the following areas such as a physical and logical security mechanism that is adapted to the needs of the company and to employee use then the procedure for managing updates and finally it needs an up to date documented system.

Information system security is often the subject of metaphors. It is often compared to a chain in the example that a system's security level is only as strong as the security level of its weakest link. All this goes to show that the issue of security must be tackled at a global level and must comprise the following elements like making users aware of security problems then the logical security, i.e. security at the data level, notably company data, applications and even operating systems and also products used in Telecommunications security such as network technologies, company servers, access networks, etc.

Data leakage happens every day when confidential business information such as customer or patient data, source code or design specifications, price lists, intellectual property and trade secrets, and forecasts and budgets in spreadsheets are leaked out. When these are leaked out it leaves the company unprotected and goes outside the jurisdiction of the corporation.

This uncontrolled data leakage puts business in a vulnerable position. Once this data is no longer within the domain, then the company is at serious risk.

When cybercriminals "cash out" or sell this data for profit it costs our organization money, damages the competitive advantage, brand, and reputation and destroys customer trust. To address this problem, we develop a model for assessing the "guilt" of agents. The distributor will "intelligently" give data to agents in order to improve the chances of detecting a guilty agent like adding the fake objects to distributed sets. At this point the distributor can assess the likelihood that the leaked data came from one or more agents, as opposed to having been independently gathered by other means. If the distributor sees enough evidence that an agent leaked data then they may stop doing business with him, or may initiate legal proceedings. Mainly it has one constraints and one objective. The Distributor's constraint satisfies the agent, by providing number of object they request that satisfy their conditions.

II. WATERMARKING THE DATA

A **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.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)
International Conference on Information Systems and Computing (ICISC-2013), INDIA.

It provides an effective watermarking technique geared for the relational data. This technique ensures that some bit positions of some of the attributes of some of the tuples contain specific values. The tuples, attributes within a tuple, bit positions in an attribute, and specific bit values are all algorithmically determined under the control of a private key known only to the owner of the data. This bit pattern constitutes the watermark. Only if one has access to the private key then it is possible to detect the watermark with some high probability.

Detecting the watermark neither requires access to the original data or the watermark. The watermark can be detected even in a small subset of a watermarked relation as long as the sample contains some of the marks. Protection of these assets is usually based upon the insertion of digital watermarks into the data. The watermarking software introduces small errors into the object being watermarked. These intentional errors are called marks and all the marks together constitute the watermark. The marks must not have a significant impact on the usefulness of the data and they should be placed in such a way that a malicious user cannot destroy them without making the data less useful.

In Digital Media such as video, audio, images, text the information are easily copied and easily distributed via the web. While sharing secured information as provided some traditional data like Stock market data, Consumer Behavior data (Wal-Mart), Power Consumption data, Weather data the Database outsourcing is a common practice. So the Watermarking provides an effective means for proof of authorship by signature and the data as the same object and also it provides an effective means of tamper proofing by integrity information is used and embedded in the data.

III. Need For Data Allocation

Information systems are generally defined by the company's data and the material and software resources that allow a company to store the data and circulate this data. Information systems are essential to companies and must be protected as highest priority. Organization securities generally consists in ensuring that an organization's material and software resources are used only for their intended purposes and also it needs to provide Information privacy, or data privacy and that is the relationship between collection and dissemination of data, technology, the public expectation of privacy, and the legal and political issues surrounding them.

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.

A data breach occurs in the Organizational information systems at the time of 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.

IV. Introduction To Cloud Computing

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. And, to the user, the technology and infrastructure behind the cloud is invisible. It isn't apparent (and, in most cases doesn't matter)whether cloud services are based on HTTP, HTML, XML, Java script, or other specific technologies.

From Google's perspective, there are six key properties of cloud computing:

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)
International Conference on Information Systems and Computing (ICISC-2013), INDIA.

- *Cloud Computing is user-centric.* Once you as a user are connected to the cloud, whatever is stored there -- documents, messages, images, applications, whatever – becomes yours. In addition, not only is the data yours, but you can also share it with others. In effect, any device that accesses your data in the cloud also becomes yours.
- *Cloud computing is task-centric.* Instead of focusing on the application and what it can do, the focus is on what you need done and how the application can do it for you., Traditional applications—word processing, spreadsheets, email, and so on – are becoming less important than the documents they create.
- *Cloud computing is powerful.* Connecting hundreds or thousands of computers together in a cloud creates a wealth of computing power impossible with a single desktop PC.
- *Cloud computing is accessible.* Because data is stored in the cloud, users can instantly retrieve more information from multiple repositories. You're not limited to a single source of data, as you are with a desktop PC.
- *Cloud computing is intelligent.* With all the various data stored on the computers in the cloud, data mining and analysis are necessary to access that information in an intelligent manner.
- *Cloud computing is programmable.* Many of the tasks necessary with cloud computing must be automated. For example, to protect the integrity of the data, information stored on a single computer in the cloud must be replicated on other computers in the cloud. If that one computer goes offline, the cloud's programming automatically redistributes that computer's data to a new computers in the cloud.

Computing in the cloud may provide additional infrastructure and flexibility.

4.1 Databases in cloud computing environment

In the past, a large database had to be housed onsite, typically on a large server. That limited database access to users either located in the same physical location or connected to the company's internal database and excluded, in most instances, traveling workers and users in remote offices.

Today, thanks to cloud computing technology, the underlying data of a database can be stored in the cloud, on collections of web server instead of housed in a single physical location.

This enables users both inside and outside the company to access the same data, day or night, which increases the usefulness of the data. It's a way to make data universal.

V. RELATED WORK

Reference Paper 1: Rights Protection is provided for Relational Data

Radu Sion,Mikhail Atallah,and Sunil Prabhakar focus on providing the rights protection for relational database using watermarking technology

Rights protection for relational data is of ever increasing interest, especially considering areas where sensitive, valuable content is to be outsourced. It handles data security through watermarking in the framework of numeric relational data and instead of primary key it uses the most significant bits of the normalized data set. Mainly, it divides the data set into partitions using markers and then varies the partition statistics to hide watermark bits.

It proposes a watermark embedding algorithm such that it consists of Sorting, Partitioning used for marker location and bit embedding watermark bits are embedded in the number set so as to provide a right protection to the data that are stored into it the relational database.

Then it also develops a watermark detection algorithm such that it consists of Sorting, Partitioning used for marker location and bit detection algorithm such that it consists of Sorting, Partitioning used for marker location and bit detection technique at the time of retrieving data from the database in its client side.

The major drawback is that it should not deal on the area of data security through watermarking in the framework of nonnumeric encoding domains in this relational database.

Reference paper 2: Watermarking Technique for Multimedia Data

Hartung and Kuttur focus on the Multimedia watermarking technology that has evolved very quickly during the last few years. A recent proliferation and success of the Internet, together with availability of relatively inexpensive digital recording and storage devices has created an environment in which it became very easy to obtain, replicate and distribute digital content (music, video, and image) publishing industries, because technologies or techniques that could be used to protect intellectual property rights for digital media, and prevent unauthorized copying did not exist.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)
International Conference on Information Systems and Computing (ICISC-2013), INDIA.

While the encryption technologies can be used to prevent unauthorized access to digital content, it is clear that encryption has its limitations in protecting intellectual property rights once content is decrypted, and there's nothing to prevent an authorized user form illegally replicating digital content. Some other technology was obviously needed to help establish and prove ownership rights, track content usage, ensure authorized access, facilitate content authentication and prevent illegal replication.

A digital watermark is information that is imperceptibly and robustly embedded in the host data such that it cannot be removed. A watermark typically contains information about the origin, status, or recipient of the host data. It provides the requirements and all the related applications for watermarking is reviewed. The application includes copyright protection, data monitoring, and data tracking. Robustness and security aspects are also discussed in specific data source.

Finally, a few remarks are made about the state of the art and possible future developments in watermarking technology.

Reference Paper 3: Achieving K-Anonymity Privacy Protection

Latanya Sweeney deals about generalization and suppression techniques to safeguard the data from the data distributors using k-anonymity privacy protection. The data in the system is analyzed for generalization like replacing or recoding a value with a less specific but semantically consistent values and suppression involves not releasing a value at all. It achieves that the released records adhere to k-anonymity, which means each released record has at least (k-1) other records in the release whose values are indistinct over those fields that appear in external data. So, k-anonymity provides privacy protection by guaranteeing that each released record will relate to at least k individuals even if the records are directly linked to external information.

The preferred Minimal Generalization Algorithm (MinGen), which is a theoretical algorithm presented herein, combines these techniques to provide k-anonymity protection with minimal distortion. The real world algorithms Datafly and m-Argus are compared to MinGen. Both Datafly and m-Argus use heuristics to make approximations, and so, they do not always yield optimal results. It is shown that Datafly can over distort data and m-Argus can additionally fail to provide adequate protection to the stored records.

It mainly focused towards suppression technique which is nothing but it should not provide the data to the user.

The major drawback in this system is that, there is no clear explanation on, how the data is going to be secured in suppression technique. The next issue is, by considering the data when it is not semantically linked then the suppression technique should not be effective.

Reference Paper 4: Watermarking the relational databases

Rakesh Agrawal and Jerry Kiernan focus on watermarking the relational databases. It suggested that watermark can be applied to any database relation having attributes which are such that changes in a few of their values do not affect the applications.

They enunciate the need for watermarking database relations to deter their piracy, identify the unique characteristics of relational data which pose new challenges for watermarking, and provide desirable properties of a watermarking system for relational data. A watermark can be applied to any database relation having attributes which are such that changes in a few of their values do not affect the applications. Then they present an effective watermarking technique geared for relational data. This technique ensures that some bit positions of some of the attributes of some of the tuples contain specific values. The tuples, attributes within a tuple, bit positions in an attribute, and specific bit values are all Algorithmically determined under the control of a private key that only known by the owner of the data. This bit pattern constitutes the watermark. Only if one has access to the private key can the watermark be detected with high probability. Detecting the watermark neither requires access to the original data, nor the watermark. The watermark can be detected even in a small subset of a watermarked relation as long as the sample contains some of the marks. Our extensive analysis shows that the proposed technique is robust against various forms of malicious attacks and updates to the data. Using an implementation running on DB2, They also show that the performance of the algorithms allows for their use in real world applications.

The major flaw is that, it should not explain how the knowledge about the schema and watermark will be given to the other user and not sure, how the owner will identify the criticality of the data to be changed.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)
International Conference on Information Systems and Computing (ICISC-2013), INDIA.

Reference Paper 5: Lineage Tracing General Data warehouse Transformations

Yingwei Cui and Jennifer Widom focus on transformation or modification of data happening automatically due to mining of data or while storing the data in the warehouse.

In a warehousing environment, the *data lineage* problem is that of tracing warehouse data items back to the original source items from which they were derived. It formally defines the lineage tracing problem in the presence of general data warehouse transformations, and they present algorithms for lineage tracing in this environment. The tracing procedures takes advantage of known structure or properties of transformations when present, but also work in the absence of such information. Their results can be used as the basis for a lineage tracing tool in a general warehousing setting, and also can guide the design of data warehouses that enable efficient lineage tracing.

The major drawback is that it should not focus on the latest tools which will solve this kind of problem automatically and there is no clear explanation is given at its security part of this technique.

Reference Paper 6: Databases in the Cloud: a Work in Progress

Edward P. Holden, Jai W. Kang, Dianne P. Bills, Mukhtar Ilyassov focus on trial of using cloud computing in the delivery of the Database Architecture and Implementation in the cloud.

It describes a curricular initiative in cloud computing intended to keep our information technology curriculum at the forefront of technology. Currently, IT degrees offer extensive database concentrations at both the undergraduate and graduate levels. Supporting this curriculum requires extensive lab facilities where students can experiment with different aspects of database architecture, implementation, and administration. A *disruptive technology* is defined as a new, and often an initially less capable technological solution, that displaces an existing technology because it is lower in cost. Cloud computing fits this definition in that it is poised to replace the traditional model of purchased-software on locally maintained hardware platforms.

From this perspective in academic, cloud computing is utilizing scalable virtual computing resources, provided by vendors as a service over the Internet, to support the requirements of a specific set of computing curricula without the need for local infrastructure investment.

Cloud computing is the use of *virtual computing technology* that is scalable to a given application's specific requirements, without local investment in extensive infrastructure, because the computing resources are provided by various vendors as a service over the Internet.

VI. CONCLUSION

The basic approaches for leakage identification system in various areas and there by proposing a multi-angle approach in handling the situational issues were all studied in detailed.

When the occurrence of handover sensitive data takes place it should always watermarks each object so that it could able to trace its origins with absolute certainty, however certain data cannot admit watermarks then it is possible to assess the likelihood that an agent is responsible for a leak, based on the overlap of the data with the leaked data and also based on the probability that objects can be guessed by any other methodologies.

REFERENCES

[1] R. Sion, M. Atallah, and S. Prabhakar, "Rights Protection forRelational Data," Proc. ACM SIGMOD, pp. 98-109, 2003.

[2] R.Agrawal and J. Kiernan, "Watermarking relational databases". In VLDB '02: Proceedings of the 28th international conference on Very Large Data Bases, pages 155–166. VLDB Endowment, 2002.

[3] Hartung and kutter,"Watermarking technique for multimedia data"2003.

[4] Y. Cui and J. Widom. " Lineage tracing for general data warehouse transformations". In The VLDB Journal, pages 471–480, 2001.

[5] L. Sweeney, "Achieving K-Anonymity Privacy Protection Using Generalization And Suppression," http://en.scientificcommons.org/43196131, 2002.

[6] Edward P. Holden, Jai W. Kang, Geoffrey R. Anderson, Dianne P. Bills, Databases in the Cloud: A Work in Progress,2012.

[7] Michael Miller, "Cloud Computing" Web-Based Applications that change the way you work and Collaborate Online,Pearson Education,2012.