Nano World Journal

https://doi.org/10.17756/nwj.2023-s4-083

Research Article

Open Access

A Novel and Efficient Detection of Data Leak with Effective Data Allocation Strategies on Cloud

S. Saisree and V. Balaji*

Department of Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad, Telangana, India

***Correspondence to:**
V. Balaji
Department of Computer Science and Engineering,
Vardhaman College of Engineering,
Hyderabad, Telangana, India.
E-mail: v_balaji@vardhaman.org

Received: September 19, 2023
Accepted: December 04, 2023
Published: December 07, 2023

Citation: Saisree S, Balaji V. 2023. A Novel and Efficient Detection of Data Leak with Effective Data Allocation Strategies on Cloud. *NanoWorld J* 9(S4): S494-S499.

Published by United Scientific Group

Abstract

With the growth of internet technologies data handling and sharing is inevitable and this is achieved by cloud computing. When a data outsourcer distributes sensitive data to the set of their subscribed and authentic agents, it is always possible that one of the authentic agents may leak the data to an unauthentic third party, this causes the sensitive data leak. Cloud computing plays an important role in data storage and sharing among agents. Sensitive data leak may cause illegitimate data usage among third party agents. When the data outsourcer comes to know that the sensitive data is leaked by witnessing data in the web or other platforms, he must be able to detect the guilt agent, thus he can be able to stop sharing further data and take actions. The proposed work achieves the objective of guilt agent detection and data leak detection by introducing fake data tuples/objects and the method of probability of distribution. The fake object is allocated on runtime dynamically and unique for every agent. The uniqueness of the fake object allows it to detect the agent who leaks the data. There are existing techniques like watermarking and anonymization to outsource data. Though the leak can be detected by the existing techniques the fault agent detection cannot be achieved, moreover the data is modified in some form and may affect the originality of data. This proposed work achieves both goals effectively.

Keywords

Data outsourcer, Cloud Storage, Watermarking, Perturbation, Anonymization, Data privacy, Data leak, Sensitive data, Probability detection

Introduction

In today's information world, the transfer of data from one individual or institution to the other is inevitable. The data outsourced from one source to the other must be handled securely to prevent any leakage. The data if it is sensitive, such as patient's information, government secure document, etc., are to be given more attention for detecting any leak. The data outsourcer is the one who gives data to another person or organization, the agents are the ones who receive data from the outsourcer. With the growth of usage of internet services, cloud computing offers more flexible and effective services for large data storage and sharing. Though the advantages of cloud computing are vast, it is highly vulnerable to security attacks. Thus, the data stored in the cloud can be provided with high security.

Data leak is unauthorized transfer of data from a legal data outsourcer to the third party illegitimately. Data leakage occurs when sensitive data such as health care data, design documents, trade details, intellectual property, or any government official data may be leaked. When data leaks, jurisdiction breach occurs as confidential data seen by many. The uncontrolled leak of information may keep business threats, as the data not in safe and secure it may lead to serious threats. The data leak may occur due to intentionally by the guilt agents or may occur accidentally.

Perturbation and animalization are the techniques in which data is modified to provide it as fewer sensitive data thus data leak may not cause serious threats. There are certain applications where the original data cannot be modified and shared, thus raising the serious issue when the data is leaked. For example, the data outsourcer creates payroll data, where the salary of the employees cannot be modified and similarly, in banking applications, where the customer data cannot be modified. There are existing works available, which handled data leak detection with the help of watermark, machine learning techniques, and some of the studies handled the data sharing with encryption and secure channel to other users/data consumers. This problem to be addressed to make the detection of data leak and if possible then guilt agent who leaked the data to take further actions.

In this proposed work, a novel approach for data leak detection and if possible untrusted agents who leaked the data is also studied. When the data is outsourced to the set of data consumers/agents, there are certain cases where the same data is seen at unauthorized places of the website, then the data outsourcer can come to know that the data is leaked. The proposed system constructs an algorithm for adding fake objects to the data in real time, while allocating themto the agents. The fake objects allocation is done in such a manner that they are unique; thus, the detection of guilt agent is made efficient. The following figure represents the proposed data leak detection framework, the data outsourcer is the one who gets a data request from agents. The data outsourcer handles the data from cloud storage.When the data outsourcer gets a data request, he allocates the data with added fake objects to the allocated data. The data allocation strategy is followed to allocate the data. Among the set of received agents, an agent may turn guilt by leaking the data to unauthorized users (Figure 1).

The proposed work fulfills the following objectives: (i) Data allocation strategies provided with fake objects creation and distribute data to the set of agents, (ii) The data allocation strategies also includes the maintaining a copy of data to detect the leakage in future, and (iii) With the use of probability distribution function for detecting the guilt agent who leaked the data with thecomputation of probability of fake objects in the leaked data.

Innovative approaches to improve storage capacity, effectiveness, and performance can result from combining nanotechnology with cloud storage allocation mechanisms.

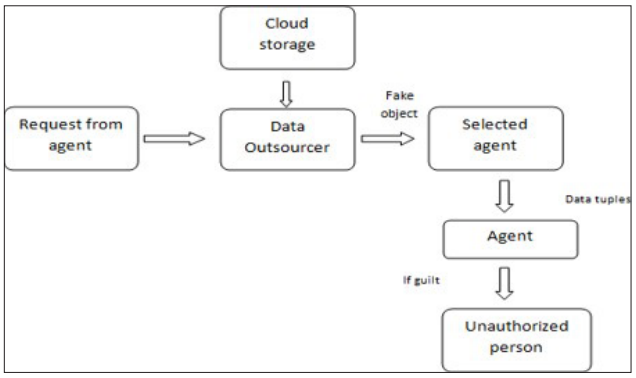


Figure 1: Overview of data leak detection.

- Nanotechnology in storage devices: Making use of nanotechnology in storage device design and production. The efficiency of cloud storage systems can be affected by nanoscale materials and structures, which can result in increased storage densities, faster data access, and lower power consumption in storage devices.
- Enhanced storage density: Cloud servers can now use nanotechnology-enabled storage media. Device storage density can be greatly increased by using nanoscale materials, such as memory cells with nanoscale dimensions. Higher capacity for cloud storage servers and more effective use of physical space may arise from this.
- Energy-efficient nanotech storage: Using nanotechnology to create cloud-based energy-efficient storage options. The creation of power-efficient storage devices may be made possible by nanomaterials. This is especially important in cloud systems, where cost-effectiveness and sustainability depend on energy efficiency.
- Integration of nanoscale computing: Including components of nanoscale computing in storage systems. By incorporating nanoscale computing into storage devices, cloud storage systems can operate more efficiently by requiring fewer large data transfers and enabling local processing capabilities.

Related work

Guilt agent detection is one of the important research projects in data sharing applications. There is much research being done on this problem. The watermarking technique and perturbation techniques are most commonly used data sharing techniques.

Watermarking is an efficient technique for sharing data among the users. This technique can handle media data such as images and videos for secure sharing. There are techniques like zero sharing and lossless watermarking that are studied by the researchers. The work [1] represents the zero watermarking for enhanced data security reasons on sharing medical images is studied. The work handled novel watermark registration and authentication, this first part of work generates features based on nearest neighbor residual grayscale and this is used for owner share, while retrieving the watermark, these features are retrieved and verified for owner share.

To solve the problem of watermarking in the bit stream domain, an anti-recompression mechanism is proposed in [2], the authors studied that the bitstream domain This works based on the bit stream is replaced with code word of numerals such as for suffix 2, level 1 replaces code 100. Similarly, the bit stream is constructed as a watermark for secure sharing of data. Watermark is embedded on the code level in macro blocks of images.

Internet of things has growing demand in this information world, the data shared among internet of things networks has the threat of leak, this is studied by authors in [3] with the use of Adaptive Feature Graph Update model. In this work, the data considered is text document, thus the pre-processing is performed with Term Frequency-Inverse Document

Frequency technique before training. The feature graph is built, and training is performed, the confidentiality of the document is measured in terms of sensitivity score of the document.

The data transformation after update of certain data is more complex to detect when it is leaked, this problem is addressed in [4] by authors using the detection of patterns. Content and sensitive data sequences are evaluated for data leaks. Content sequences are nothing, but data retrieved from network channels and comparable sampling is the method used for context aware selection.

Order preserving encryption is proposed in [5] for perturbation of data, in two works two types were adapted, one is mutable order preserving and another one is storage aware order preserving. The random perturbation scheme proposed has a proof of correctness algorithm to check the insert or delete operation correctness. This proposed technique is useful for client server applications as well as cloud-based environments.

Adaptive weighted graph walk method was proposed by authors in [6] for handling unstructured big data problems for data leak. This method is handled in three steps, first it quantifies test data for sensitive data, label propagation is implemented in handling fresh data. The author implemented a score walk algorithm to check the sensitivity.

The literature survey of the data leak detection methods discussed here infer that the traditional system like watermarking and perturbation methods need change in the data, which may affect the originality of the data, thus an effective system to handle the data leak detection without changing the data is required. The proposed work is based on adding fake objects to the existing data, this technique does not change the data, thus its sensitivity is not disturbed.

Experimentation

Proposed work

The proposed work is identifying leaks in outsourced data. The proposed work considered patient's data as it is sensitive data, the leak in data should be identified and if possible, also identify the agent who leaked it. In this proposed work, a novel method is proposed for identifying the leak detection. The proposed work is achieved with three modules and algorithms are data outsourcers allocating the data to a set of legitimate agents, the second one is detection of probability for the leaked file by considering the leaked tuples and third one is identifying the guilty agent who leaked the data. Thus, the work has three entities to perform these modules are data outsource, legitimate agent and unauthorized third party. The proposed work assumptions and needs are addressed below:

- After allocating the data with a set of fake objects to agents, the data outsourcer finds some of the same objects in an unauthorized place.
- Data outsourcer computes the likelihood probability for the leaked data, which has been leaked by one or more agents for computation and evaluation of identifying guilt agents.
- If the data outsourcer finds the guilt agent with high probability, he may stop further business and may take legal actions.
- The proposed work handles data allocation algorithms based on data request type from agents. Thisalgorithm is proposed for enhancing the identification of guilt agents who leak data to unauthorized users.
- Fake objects added to the data, which are outsourced to agents, appear like real objects, whereas the fake object details and object ID are maintained by the data outsourcer to refer in future for leak identification.
- If the probability of fake objects in the leaked data is high, then data outsourcers can identify the guilt agent with high probability.

A novel technique for detection of data leak is proposed combining allocation strategy and probability likelihood detection. Adding too much of fake objects may make the problem tedious, however, adding few objects and with a copy of fake objects in data outsourcer cloud storage can be tracked and considered for non-treatment, as in this proposed system, patient's records are considered, the fake objects are considered to be non-treated. When the number of data requests from the agents is few, this also helps in detecting the guilt agents as the number of requests is very less. Fake objects help in finding the guilt agent reasonably with high probability of identifying the guilt one correctly. The probability increases when all the fake tuples are found in the leaked data. The advantage of this system is that the agent does not know fake tuples, as it all resembles a normal data tuple.

The work involves developing a framework for detection of guild agents. Algorithms for automatic data allocation, probability likelihood detection, and algorithm for identifying guilt agents were proposed. The fake objects appear to be real object agents and cannot find any difference; however, they are completely maintained in the cloud database by the data outsourcer. Cloud server manages the entire patient data and a copy of data allocated to all agents along with the copy of fake data allocated to agents. Even when the same data is allocated to two agents the fake object is different for both agents, thus when the data leak there is a high chance of identifying the guilt agent correctly. Thus, the fake objects addition to the distributed data gives minimized overlapping of data between agents, who receivethe same data from the data outsourcer.

It is trivial to identify the guilt agents, who leak data to illegitimate persons and publish it on public platforms. The identification of such agents without modifying the original data is very difficult, as the existing system dealt with these problems vastly on watermark methods, data anonymization methods and perturbation techniques. However, these existing methods involved any or small form of changes to the original data. There are certain conditions where one cannot modify the original data such as employee's salary list, patient's data, government voter's ID list and other confidential data and more. Similarly, perturbation or anonymization cannot be

performed in certain data, where patient age in some cases doctors need to understand the proper age to identify the risk of diseases and salary cannot be anonymized with generalization methods as it needs the correct digit to pay the employees.

Figure 2 represents the overall system architecture of proposed work. The data is stored in cloud storage and accessed by the data outsourcer, who distributes data to all agents. The set of agents who receive data from outsourcers are authorized and legal agents. As shown in the figure, on allocation time, fake objects are introduced to the overall data and allocated to agents. One of the agents may leak data to an unauthorized/ illegitimate person who received data from the agent and publish it on a public platform.

Modules

The proposed work is implemented with four modules. These modules handle the overall data allocation process and identifying guilt agents.

Data request from agents

There are a set of agents {A1, A2, ..., An} under a data outsourcer, these agents are considered to be legitimate and authentic to share the data for their usage. There are two types of requests from agents are 'Random' and 'Explicit'. The random method allocates objects based on a random number of tuples, whereas Explicit allocates data based on patient name requests from the agents.

Data allocation strategy

Data allocation is performed by the data outsourcer to the set of agents upon their request as explained in the above module. The agent gets data with additional fake objects created by the data distributor during run time. In this study, as the patient's data is handled, the original records in the cloud storage contains attributes {patient_name, age, sex, mobile}. The fake object added by the data distributor is patientID. Thus, agents received data contains attributes {patient_name, age, sex, mobile, patientID}. As the patientID is the fake object introduced by the data distributor it appears to be the real object for agents, they cannot understand and identify that it is a fake object. Thus, it increases the chances of guilt agent identification. The probability of detection rate increases with the uniqueness in the fake object distributed to agents and minimizes the data overlapping problem. The algorithm 1, data distribution algorithm explains the allocation based on the given requests from agents as given below (Figure 3).

Algorithm 1: Data distribution algorithm

Input: Set of Agents A1, A2, …., An, Condition = {Random(R), Explicit(E)}, Data = {T1, T2, …., Tn}, FO = {TF1, TF2, …., TFn}

Where A1, A2, ..., An, are agents, FO - Fake object.

Step 1: Data outsourcer gets data request from agents along with condition 'E or R'.

Step 2: If (Request = E).
Allocate data according to patient name.
Else,
Allocate data randomly.

Step 3: Data outsourcer on allocation creates fake object with unique ID, here random number is generated as patientID and added to the original data.

If the distributor can create more fake objects, then the objective of the study can be improved considerably.

Step 4: Fake object created with six digits, first two digits are the random number, and the next four digits specify theagent identity.

Probability likelihood detection

The guilt agent should be detected accurately otherwise false claims may happen, to avoid this problem, the proposed method introduced unique fake object to the agents. The fake object is created in such a manner that last part of the patientID is appended with the agent ID. The leaked file is considered for evaluation and detection of guilt agents. The probability is calculated based on the matching number to records to the total number of records. The problem arises here is that if one more agent received the same tuple from data distributor, then the probability may be the same or even high for the innocent agent. Thus, the agent detection must be performed accurately with the help of a fake object. This unique object is used for the comparison and detection of leaked files and distributed data to agents, thus identifying the guilt perfectly. The algorithm for probability likelihood detection is given below.

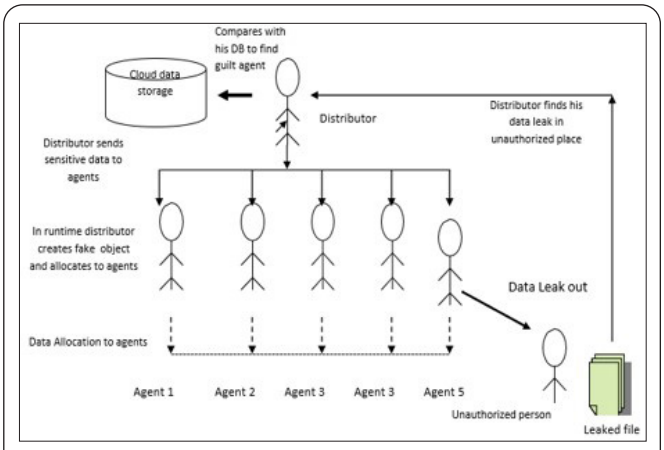


Figure 2: System architecture.

Patie...	Age	Sex	Mobile	Pati...
latha	12	F	987...	812...
kumar	12	M	787...	332...
kan...	70	M	777...	552...
kan...	23	M	989...	892...
krish	28	M	255...	312...

Figure 3: Data allocated using sample request to agent.

Algorithm 2: Probability likelihood detection

Input: Leaked file with tuple {T1, T2, …., Tn} Output: Probability (P)

Step 1: Consider a leaked file, which contains some leaked tuple, distributor compares the leaked tuple with hisdatabase, which contains allocated data for every agent.

Step 2: Find agent probability.

P = (Matched tuple/Total tuples in file)

Probability is calculated as the ratio between the number of matching tuples in leaked data for agents to the total number of leaked tuples. The implementation results for probability likelihood computation for agents for a given leaked file is show in figure 4.

Guilt agent identification

Guilt agent is identified based on the high probability for the set of agent's probability. This is computed based on the number of matching records with fake objects thus giving accurate results on detection. The algorithm 3 explains the guilt agent identification.

Algorithm 3: Guilt agent identification

Input: Agent ID, Probability (P) for agent

Step 1: Find probability of agent using Algorithm 2.
Step 2: Agent_Probability = {P1, P2, …., Pn}.
Step 3: Agent with high probability is detected as a guilt agent.

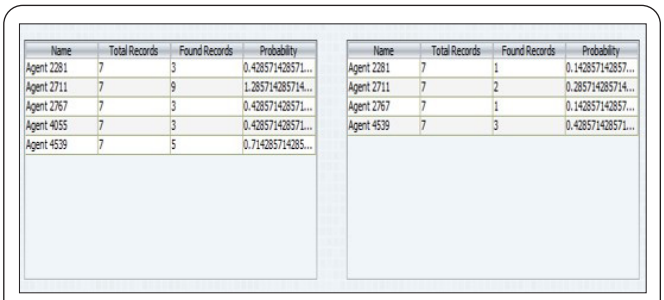
Results and Discussion

The proposed work is implemented in JDK 1.8 with an application developed in Java Swing. The cloud storage used by the data distributor for data allocation and storing the allocated tuples, MySQL database is used. The proposed method is novel based on the identification methods used. The fake object used in this implementation is unique for every agent to maximize the probability of accurate detection. In this implementation the attributes shared among the agents included a patientID column, where it is a combination of random digit and agent ID. The evaluation of guilt agent is performed before and after adding the fake objects. Figure 5 shows the results before adding fake objects. The results showed that a total of five agents and their probability of likelihood to be guilty.

Figure 6 represents probability of likelihood after adding fake objects. In these results, it is shown that there are only four agents' ID and their probability. According to the probability, there is a high probability for Agent 4539, thus it is identified who is the agent who leaked the data. This result clearly shows that the proposed novel approach eliminates the overlapping of tuples between agents and identifies the leaked data from the guilt agent.

Conclusion

In the growing information world, it is mandatory to protect the sensitive data shared by data distributors. There are techniques like watermarking available to track the information leak, however, there are certain cases where the datacannot take a watermark, in such cases a novel data leak detection technique is required to track the sensitive information leak. Unauthorized people can illegitimate and publish sensitive data on the web or another platform. The proposed work is a novel framework, which handles data allocation from cloud storage by the data distributor and tracking them till detection of leak. There are cases where the same data may be allocated to one or more agents; the proposed work also handles this trivial problem, where the fake object creation and insertion on run-time help to detect the guilt perfectly. The evaluation of results shows that the probability of likelihood detection of guilt agent is identified accurately in the proposed work.



Name	Total Records	Found Records	Probability
Agent 2281	7	3	0.428571428571...
Agent 2711	7	9	1.285714285714...
Agent 2767	7	3	0.428571428571...
Agent 4055	7	3	0.428571428571...
Agent 4539	7	5	0.714285714285...

Name	Total Records	Found Records	Probability
Agent 2281	7	1	0.142857142857...
Agent 2711	7	2	0.285714285714...
Agent 2767	7	1	0.142857142857...
Agent 4539	7	3	0.428571428571...

Figure 4: Probability likelihood detection.

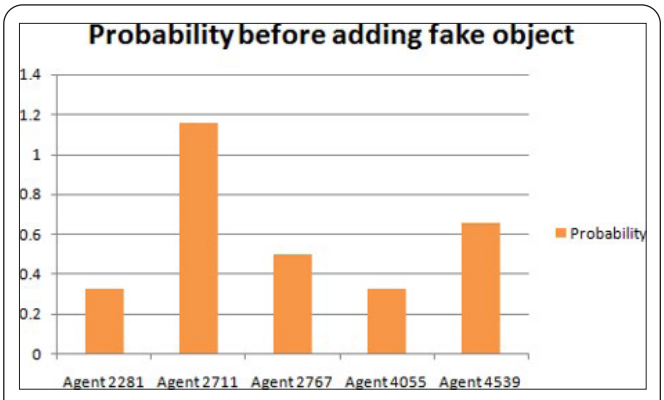


Figure 5: Results of probability of likelihood before adding fake objects.

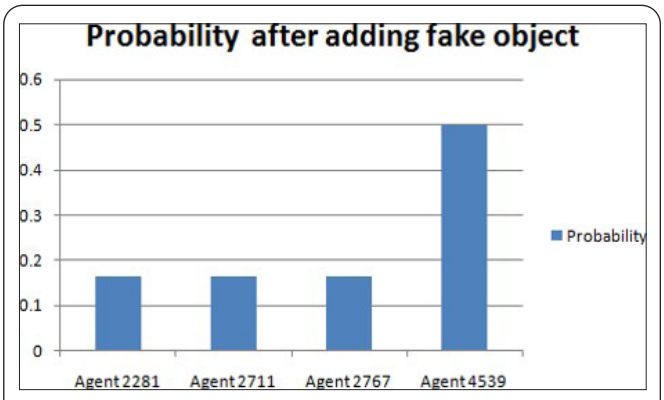


Figure 6: Results of probability of likelihood after adding fake objects.

Acknowledgements

None.

Conflict of Interest

None.

References

1. Dai Z, Lian C, He Z, Jiang H, Wang Y. 2022. A novel hybrid reversible-zero watermarking scheme to protect medical image. *IEEE Access* 10: 58005-58016. https://doi.org/10.1109/ACCESS.2022.3170030
2. Sun J, Jiang X, Liu J, Zhang F, Li C. 2020. An anti-recompression video watermarking algorithm in bitstream domain. *Tsinghua Sci Technol* 26(2): 154-162. https://doi.org/10.26599/TST.2019.9010050
3. Yu X, Qiu J, Yang X, Cong Y, Du L. 2019. An graph-based adaptive method for fast detection of transformed data leakage in IOT via WSN. *IEEE Access* 7: 137111-137121. https://doi.org/10.1109/ACCESS.2019.2942335
4. Shu X, Zhang J, Yao DD, Feng WC. 2015. Fast detection of transformed data leaks. *IEEE Trans Inf Forensics Security* 11(3): 528-542. https://doi.org/10.1109/TIFS.2015.2503271
5. Kamara MA, Li X. 2021. Random perturbation order preserving distribution encryption. *IEEE Access* 9: 165568-165575. https://doi.org/10.1109/ACCESS.2021.3130737
6. Huang X, Lu Y, Li D, Ma M. 2018. A novel mechanism for fast detection of transformed data leakage. *IEEE Access* 6: 35926-35936. https://doi.org/10.1109/ACCESS.2018.2851228