Detection Of Data Leakage In Cloud Storages

Naresh Vurukonda, Allu Venkata Dattatreya Reddy, Gutta Chiranjeevi, Kancharla Raviteja

Abstract: Leakage of sensitive data may leads to the loss of confidential and integrity. Some of the data may be leaked and found on web or untrusted users. Distributor have to take upon these situations in order to maintain data confidentiality and ensure a safe data transaction. Many small business authorities have data leak issues via internet or other means. We would like to propose a alternative methodology to implement in real world and it is different from traditional methods. Traditional methods contain "watermarking" and 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. But this also will not work if the guilt agent knows the fake objects. So the other method for getting the guilt agents is to be determined. Many methods have been in existence but every method is being override by other means using complex methodologies and by various combinations of the algorithms. These complex methods would secure much better than older ones. We are finding the agents by taking the parameters like how much time he is spending in the data, how many times he opened that file etc.... we can find the probability if the probability is more than the threshold value then we can conclude that the agent had compromised. In this model we use the previous methods knowledge to predict the agents or to over come in the solution.

Keywords : water marking,guilt agent, fake object, probability, automation

1. INTRODUCTION

In a rapidly growing digital world the sensitive data is being transferred from all parts over the globe. With the increasing concerns over the data transmission many methods are being implemented to prevent the data leakage. Early works consists of the traditional methods like Watermarking and cipher text conversions. These methodologies have been implemented on the text files and many multimedia formats and updated formats of video and audio cannot be compatible with these methods. Many specific companies and business agencies have been compatible with the watermarking methodology and used it to watermark the files transferred over networks. Watermark seems to be a prominent solution for the business model for particular time and has been overcome with time. Watermark's has been destroyed or removed using the advanced cryptography tools. Later on many methods has been to existence and some of them are able to survive a bit long like Fake object allocation, Agent Guilt model, optimization method, hashing and salted hashing. From the literature based on the previous papers related to these algorithms. We can understand that vulnerabilities can be found from all the methodologies.We could know the probability in an different way such that agent is dependent on its previous activity. All the agents are judged with their previous history of leakage and demand from an agent. Probability is calculated based on their data using Naïve Bayes or related probability algorithm to gain probability. The value is compared to a threshold level of risk and decided by the AI system to allow or not.

- *Naresh Vurukonda, Assistant professor, Koneru lakshmaiah educational institute, Guntur, India, 9908109980, nareshvurukonda@kluniversity.in*
- *Allu venkata Dattatreya Reddy, student, Koneru lakshmaiah educational institute, Guntur, India, 9533650272, dattatreya.allu.4370@gmail.com*
- *Gutta Chiranjeevi, student, Koneru lakshmaiah educational institute, Guntur, India, 9491559182, gchitanjeevi1999@gmail.com*
- *Kancharla Raviteja, student, Koneru lakshmaiah educational institute, Guntur, India, 9701329957, kancharlaraviteja1999@gmail.com*

2. RELATED WORK

The main objective of this project is to find the guilt agents means the agents that leaks the data to the third party users for some financial uses or for some other activity. Actually using the fake objects and the watermarking methods which are use earlier for finding the guilt agent are very old methods. We propose a new method for finding the guilt agents based on the number of times agent access the data and the time duration agent access the data. For this approach we have a designed a flow at which we find the guilt agent even more simple and fast when compared to remaining approaches. For finding the details of the agents like how much time agent is using and accessing the files we have different approaches and also we can design some algorithms, but it takes lot of time and it will not be accurate. So there are some online sites for doing the same purpose in a very accurate way. We have taken sales handy website for this purpose it is meant for the tracking purpose of the files and emails so we have used it. We should upload the data we want to share to the agents and have to generate the link for the following data. The link will be shared to the agents in any of the existing methods that you prefer. After that you can monitor the details of the agents in the websites in your account. The data that was present in the site have to be extracted for the further purpose for that we used automation and create a bot for automatically extracting the data out of the website without any human work.

After the data is extracted the next is to calculate the probability based on the time the agent is accessed to the data and the average time the data is accessed with machine learning and for analyzing we used R programming. We will decide a threshold value for every particular data and by using this we will calculate the probability for the agent to be guilty. In this analysis if the probability of the agent to be guilty is high then we mark that particular agent as guilty and we will not forward the data any more to that particular agent.

3. WATERMARKING:

Watermarking methodology being implemented on the text files for data leakage detection. The watermark is applied on the various parts of the files and later sends to the requested Agent. On finding the unique watermark at a unauthorized person or a firm can be considered as data

leak situation. Watermarks can be of the text, audio and image types. These can be Watermarked using particular appropriate methods. General text files can be watermarked with regular text in background. Images can be watermarked using the Cocktail Watermarking and Robust Watermarking methodologies. Cocktail watermarking can be used to alter the image using the Wavelet coefficients in particular areas of images.

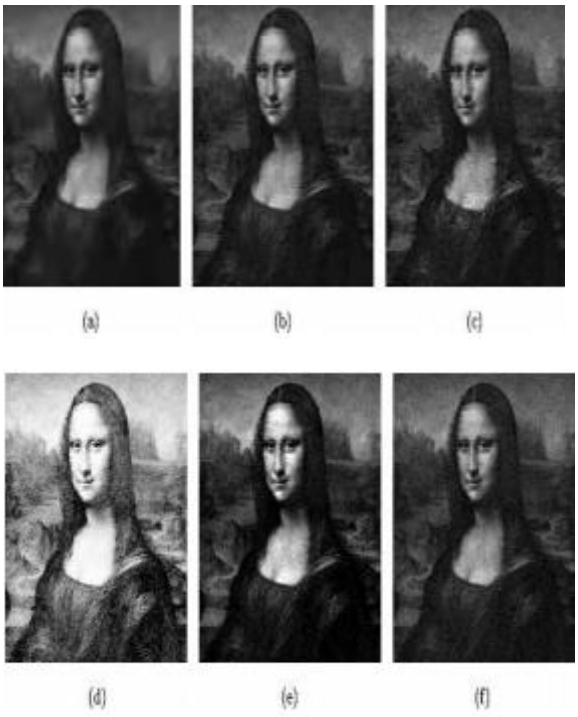


Fig 1.

Similarly images can be also altered using the robust watermarking. Results will be similar as they are images the difference cannot be found easily by a glance at it.

4. GUILT AGENTS:

In real time scenario the distributors sends the data to the agents. Agents may be of an intermediate passage or they could be of any agency requesting data to analyze it. Hospitals may act as distributor and private vendors are termed as agents. Data may be leaked from agents too, for such conditions the data will be tracked using different algorithms. These leaked agents can be termed as the guilty agents. Data leakage can be of different types. The leaker could collect data from one single agent or multiple agents so it could cause anonymity. Doing all these kind of things would lead distributor to confuse and reduce in probability so for these problems solution is needed to be found.

4.1. AGENT GUILT MODEL:

This model calculates the probability that any agent is leaking data or not. Only individual data may be collected and probability could be seen. Pr {gi/s} is the probability of an guilty agent this can be seen and agent could be blocked away from data leakage. Individually an data could be sent through each route and analytics could be seen for agent usage so while the data is being traveled network algorithms could monitor the leakage. Our end system would develop a system to check the probability of leakage. total event could be seen and individual leakage could be seen so the probability could be obtained. for example probability of one fake mail among 10 mails is 0.1.

4.2. DATA ALLOCATION PROBLEM:

The distributor should allocate the data to agents in a monitored manner such that no leakage would be detected and if any it should be noticed. For this the Distributor sends the data in a mixed format. The data may be of duplicate or original data or complete fake data. Agent cannot determine these only the methods like hashing or salted hashing can be seen. In other case the data contains any fake data. The fake data is mixed and used to determine the guilty agent incase of any malicious activity is noted.

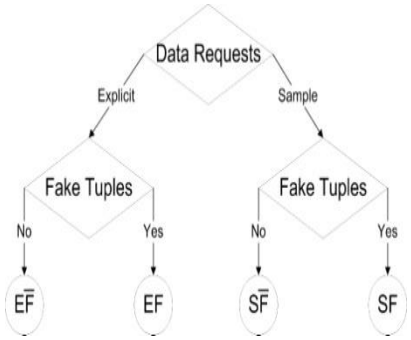


Fig 2.

5. FAKE OBJECT:

Fake objects cannot be recognized easily by any third party or agents. These fake objects could be of unnecessary data in any format. Randomly the fake objects will be allocated to data sets and in case of leak the information is verified using the fake objects. It is an updated version of watermarking. All these kinds of data will be monitored over the web and will be checked all the way for any leakages.

6. OPTIMIZATION PROBLEM AND

APPROXIMATION:

The optimization is based on allocation of data by a distributor to the agents to satisfy the agent's requests. Agents may request n number of objects for work purpose, Distributor should send the data and detect the data leakages in the process. Agents requests data as per their requirements and distributor should not deny the data requests and process the request with fake objects as mentioned in [7].

Input: $R_1, \ldots, R_n, cond_1, \ldots, cond_n, b_1, \ldots, b_n, B$
Output: $R_1, \ldots, R_n, F_1, \ldots, F_n$

$\mathbf{R} \leftarrow \emptyset$ ▷ Agents that can receive fake objects
for $i = 1, \ldots, n$ **do**
  if $b_i > 0$ **then**
    $\mathbf{R} \leftarrow \mathbf{R} \cup \{i\}$
  $F_i \leftarrow \emptyset$ ▷ Set of fake objects given to agent U_i
while $B > 0$ **do**
  $i \leftarrow$ SELECTAGENTATRANDOM$(\mathbf{R}, R_1, \ldots, R_n)$
  $f \leftarrow$ CREATEFAKEOBJECT$(R_i, F_i, cond_i)$
  $R_i \leftarrow R_i \cup \{f\}$
  $F_i \leftarrow F_i \cup \{f\}$
  $b_i \leftarrow b_i - 1$
  if $b_i = 0$ **then**
    $\mathbf{R} \leftarrow \mathbf{R} \backslash \{R_i\}$
  $B \leftarrow B - 1$

7.RPA

Robotic Process Automation is emerged in early 2000. It mostly relies on the “Artificial intelligence” and “Screen scraping technology”. By the end of the decade, with parallel to “Artificial Intelligence” and “Machine Learning”, RPA has an rapid growth and it reduces the mundane tasks. UIpath, Automation Anywhere, Blue Prism are widely used frameworks of RPA.

8. PROPOSED FRAMEWORK

Our goal is to detect, when the distributor’s sensitive data has been leaked by agents, and if possible to identify the agent that leaked the data. There are several techniques for detecting leakage of a set of objects or records. In this section we develop a model for assessing the “guilt” of agents. There are different algorithms for distributing objects to agents, but we are distributing the data by creating a link and we will track the analysis of the link. After that we will import the data using automation and we have calculated the probability of the agent to be guilty. By this analysis we have removed the access to the agents based on the above analysis.

8.1. Tracking details:

There are several methods for tracking the agent’s details (what action he is performing or how much time he spent on the data) but we have chosen sales handy site for tracking the details of all the agents.
In this a is link is created for the data that we want to share to the agents. Then the link that was created will be shared to the agents. From now the actual task begins, the tracking details will be automatically updated in the above site our task is to save the data into the excel sheet with the help of automation.
If the agent access the data that will be updated automatically and accurately. If we use any algorithm for tracking the data, it will be difficult to maintain the whole process, so using some online sites is easy and we can do this accurately.

Name	Link Name	Time Spent	Viewed	Visit Detail	Time	Action
Gdh@gmail.com	resume	00:01:14	100.00 %		17 hours ago	
Raviteja@gmail.com	resume	00:00:10	50.00 %		17 hours ago	
naresh@gmail.com	resume	00:10:40	100.00 %		17 hours ago	

Fig 3.

These are the details that are visible for us in the tracking site. The overall time that was spent by each individual will be updated time to time. How much content was seen by the agent and how much time ago it was seen all the information will be able to tractable.
The total time that was taken by the agents and the average time that was taken will displayed in the website this makes us comfortable for calculating the probability of the agents to be guilty. We can also track the individual performance like if the agent shares the file any other person then updated in the site and if the agent downloads the file then the information related to the agent also will be updated.

Fig 4.

As above displayed all the information that related to the shared document will be tracked and by using the above information we will able to calculate the probability.

8.2. Exporting the data:

The data that was recorded should be exported for further analysis. We can manually export data, but it is difficult to export manually. So by using so modern technology called automation we can export the data automatically. The prototype for this

Process is displayed bellow.

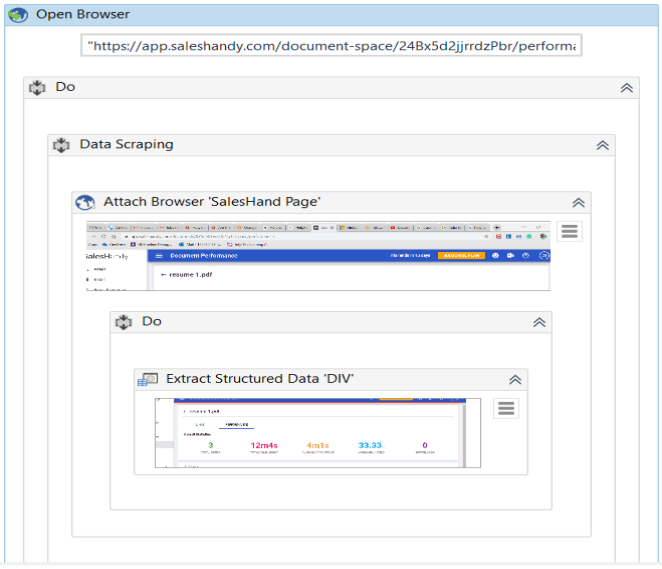


Fig 5.

By using above model we have imported the data into the excel sheet. In automation we have created a bot that automatically imports all the data into the excel sheet when we run the program. Automation makes the work simple and reduces the effort of the human. It is equal to programming language and it is more than that. After writing the flow chart in the uipath for collecting the data into the excel sheet then simply run the flowchart, within the

fraction of seconds the data will be collected into required format.

Name	Link Name	Time Spen	Viewed	Visit Detail	Time
naresh@gı	resume	0:10:40	100.00%		3 hours ago
Raviteja@	resume	0:00:10	50.00%		3 hours ago
Gdh@gma	resume	0:01:14	100.00%		3 hours ago

Fig 6.

8.3. Calculating probability:

This is the required excel sheet for calculating the probability for finding the guilt agents. There are several existing methods for calculating the probability for finding the guilt agents but his is a different approach for finding the guilt agents. The agent's data that was existing with us will be analyzed using R programming. In this first we should import the data into the R module and after that we should write some sort of code to analyze the data in R. A threshold value should be taken based on the type of data and size of the data. It differs from data to data, that should be selected by us. By using this threshold value we can calculate the guilt agents. The probability is calculated based on the overall time used by the agents to the individual time used by them. If the probability of the particular agent is more than the threshold value then we can assume that particular agent is guilt agent. We can plot different graphs for the analysis of the agents to be guilty. Based on the probability and the data shown by the graphs the agent to be guilty will be determined. Actually the agent may have a chance to be guilty but in this method we will be able to find the guilt agent even faster when compared to remaining methods. It is very easy to find the guilty agents.

9. CODE:

```
library(ggplot2)
thresold<-300
prob<-sum(wek$seconds)/thresold
m<-c("Agent1","Agent2","Agent3")
h1<-wek$seconds[1]
h2<-wek$seconds[2]
h3<-wek$seconds[3]
H<-
c(h1/sum(wek$seconds),h2/sum(wek$seconds),h3/sum(we
k$seconds))
barplot(H,names.arg=m,xlab="Agents",ylab="probability")
```

The above code calculates the probability of the agent to be guilty and the graph determines the agent to be guilty, by seeing the graph we can say the agent is guilty or not.

10. EXPERIMENTAL RESULTS

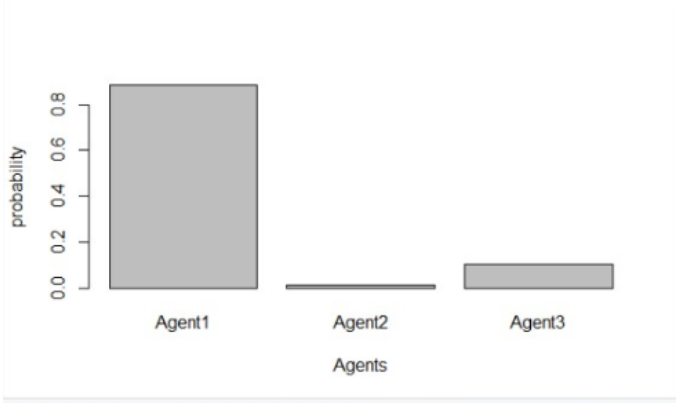
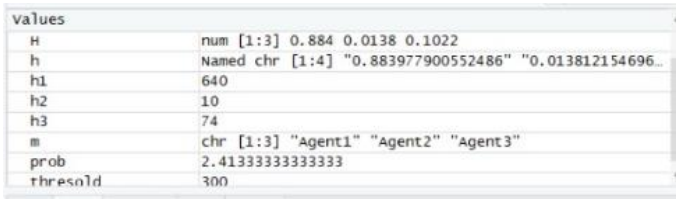


Fig 7.



Values	
H	num [1:3] 0.884 0.0138 0.1022
h	Named chr [1:4] "0.883977900552486" "0.013812154696…
h1	640
h2	10
h3	74
m	chr [1:3] "Agent1" "Agent2" "Agent3"
prob	2.41333333333333
thresold	300

Fig 8.

Fig 9.

Here there are three agents Agent1, Agent2, Agent3 suppose we have taken threshold of 0.3 based on the content shared. Based on the analysis of the agent probability of an individual agent to be guilty is calculated. Now see the bar chat in fig 7 Agent1 having probability greater than 0.3 then we consider him as a guilty agent.

11. CONCLUSION AND FUTURE WORK

In any kind of business enterprise there are some people who leak the data to third party users. To overcome this there are so many methods base on the confidentiality of the data. In some papers they used some cryptographic algorithms for encrypting the data and to provide security. Some used fake objects. Keeping all this aside the model that we have developed will be very perfect in tracking the guilt agents when compared to other methods. This method tracks the guilt agents very fast when compared to other methods. The probability that we calculate based on the time will give us the best solution for what we are looking.

REFERENCES

[1] Reddy, G. Venkatakoti et al. "A review on active data access control for multi-authority cloud storage systems with users." 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC) (2017): 262-266.

[2] Vurukonda, Naresh & Rao, Dr. B.Thirumala. (2016). A Study on Data Storage Security Issues in Cloud Computing. Procedia Computer Science. 92. 128-135. 10.1016/j.procs.2016.07.335.
[3] Abbas, Assad et al. "A cloud based health insurance plan recommendation system: A user centered approach." Future Generation Comp. Syst. 43-44 (2015): 99-109.
[4] P. Mell, T. Grance, The NIST definition of cloud computing (draft), NIST Special Publ. 800 (145) (2011) 7.
[5] Vurukonda, Naresh & Rao, Dr. B.Thirumala. (2016). A Study on Data Storage Security Issues in Cloud Computing. Procedia Computer Science. 92. 128-135. 10.1016/j.procs.2016.07.335.
[6] Chavan, Jaymala and Priyanka Desai. "Relational Data Leakage Detection using Fake Object and Allocation Strategies." (2013).
[7] Papadimitriou, Panagiotis and Hector Garcia-Molina. "A Model for Data Leakage Detection." 2009 IEEE 25th International Conference on Data Engineering (2009): 1307-1310.
[8] Wakhare Yashwant R & B. M. Patil, "Data Leakage Detection with K-Anonymit Algorithm"
[9] DATA LEAK DETECTION Research article by Ms. N. Bangar Anjali1, Ms. P. Rokade Geetanjali2, Ms. Patil Shivlila3, Ms. R. Shetkar Swati4, Prof. N B Kadu5 from "ijsmc".
[10] Periyasamy, A R. Pon and E. Thenmozhi. "Data Leakage Detection and Data Prevention Using Algorithm." (2017).
[11] Review Paper on Dynamic Mechanisms of Data Leakage Detection and Prevention by shivkumar tuppada, Muneswar M M S, Dr. Rajasekhar patil.
[12] Guevara, César et al. "Data leakage detection algorithm based on task sequences and probabilities." Knowl.-Based Syst. 120 (2017): 236-246.
[13] Rajasekaran, M & Gupta, Amisha & Sharma, Padmini. (2018). Data Leakage Prevention and Detection System. International Journal of Engineering & Technology. 7. 366. 10.14419/ijet.v7i3.12.16108.