International Journal of Engineering & Technology, 7 (2.31) (2018) 101-105

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Securing cloud by mitigating insider data theft attacks with decoy technology using Hadoop

K. Vamsi Krishna[1*], V. Srikanth[2]

[1]Master's Student, Dept. of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation (KLEF), Green Fields, Vaddeswaram, Guntur Dt., Andhra Pradesh, India.
[2]Professor, Dept. of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation (KLEF), Green Fields, Vaddeswaram, Guntur Dt., Andhra Pradesh, India.
**Corresponding author E-mail:162031004@kluniversity.in*

Abstract

Cloud Computing has been intrinsically changing the way we utilize computers to keep and retrieve our personal & business data. With the advent of this emerging paradigm of computing, it arises the new security challenges. Existent cryptographic data security techniques i.e., encryption deteriorated in preventing data theft attacks once the key is compromised, especially those perpetrated by insiders. Cloud Security Alliance reckoned this threat as a significant danger of Cloud Computing. Although the majority of Cloud users are very much known of this risk, they are leftover with the only choice of trusting the cloud service provider, regards to their data protection. In this paper, we propose an alternate way to secure data on the cloud which is more efficient and secure by the concoction of user profile mapping using Hadoop framework and offensive decoy technology.

Keywords: *Data security, data access, malicious insider, decoy, intrusion detection, fog computing, user profiling, hadoop, cluster computing, multi-clouds, machine learning, naïve bayes, big data processing.*

1. Introduction

Nowadays the need of storing data is increasing day-by-day, numerous people are moving towards digitization for having a hassle-free life. Digitization helped people to store their data in order to avoid failure to recall scenarios and of course for easy access, but this demand to store more amount of such data, devices have become incapable which conceived the idea of Cloud.

Cloud Computing is an emerging paradigm originated from Distributed Computing, which [1] endows on-demand, ubiquitarian, and convenient network access to a shared pool of configurable resources like as network, server, data storage, and applications shared among users on a subscription basis as services with minimal management effort. Typically, where the data and services reside in massively scalable data centers in the cloud and can be accessed from any of the connected devices over the internet. Many of the small and medium businesses (SMBs), especially start-ups, are choosing for offshoring their computational infrastructure to the Cloud for better operational efficiency [2].

As the span of Internet builds step by step and the adoption of Cloud Computing and Internet of Things(IoT) merge will be foreseen as disruptive and as an enabler of a large number of application scenarios, this usage is expected to be more and more pervasive, making them important components of the future internet [3].

Cloud has its benefits such as flexibility, scalability, efficiency, multi-portability, etc., it's market growth crossed beyond the expected limit. End users will get more benefits with the uninterrupted services. Also, services provided with low cost as well as overheads of reduced maintenance. But the recent attacks raised various questions in cloud security, perhaps the most serious of which data theft attacks is [2]. This has been considered as one of the major threat reported by Cloud Security Alliance [4].

Despite the fact that efforts are been taken to secure the cloud, there are still loopholes in it which is confining the users from the cloud. The exacerbation of let alone control over of the cloud providers' authentication, authorization, & audit controls [2] and lack of transparency leads to this threat.

Insider Threat stated by CERT as "A malicious insider threat to an organization is a current or former employee, contractor, or other business partner who has or had authorized access to an organization's network, system, or data and intentionally exceeded or misused that access in a manner that negatively affected the confidentiality, integrity, or availability of the organization's information or information systems." [5]. According to a study conducted by IBM, 31.5 % of data breaches originated from malicious insiders and another 23.5 % are from inadvertent actions [6]. There are three distinct types of insiders that can pose a threat to cloud security are:

- Malicious Insider - This type of insider threat is likely the most difficult to face, and the threat they pose is not easily mitigated by more stringent protocols or advanced information security training. This leads external hacker to make use of stolen credentials to login into the network; once in, they have free rein to roam around unfettered.
- Negligent/Unknowledgeable Employee – These employees can inadvertently compromise the security and safety of a cloud network.
- 3rd Party Contractor - Whether it's as simple as the maintenance company contracted, or the lab a practice outsources testing too like as the negligent or unknowledgeable employee, third-party contractors provide another opportunity for malicious hackers to compromise cloud service provider's network security.

Most recently, Tele-Communication Company Verizon breach has exposed 14 million US customers account details and PINs. Yahoo! was reported that, hackers stole personal information from 500 million Accounts, including Dates of Birth, security question & answers, and hashed passwords which are used to verify the users' identity [7]. A data breach of Apple's cloud suite iCloud revealed hundreds of celebrities' personal pictures, which was a major invasion of privacy of their subjects [8]. Another data attack is of passport of USA's first lady Michelle Obama has been posted on the web [9].

With this, it is clear that data residing on the cloud is not safe from intrusion. To overcome these threats more sophisticated mechanisms of handling authenticity of users were required.

2. Related work & literature survey

Over the years with numerous proposals, Cloud Security has been focused on broad research in on the methods of avoiding unauthorized and illegitimate access to data through the development of skeptical access control mechanisms and encryption. However, with the levels of assurance as people desired these mechanisms still got failed to prevent data thefts.

Cryptography may be seen at first sight as the answer for data confidentiality in the cloud. Although, data in IaaS cloud environments have to be handled by the applications that run on the users' Virtual Machines (VM), which normally cannot happen if the data is encrypted. There has been some talk about fully homomorphic encryption (FHE) as an answer for this issue, since this permits certain operations, in addition, to be executed over encrypted data. Although, at the present time performance of FHE makes this infeasible.

The problem with this is that with the users' access the data stored in the cloud using the users' credentials, with the above mechanisms in which the user credentials and cryptographic keys can be retrieved, it can posit severe security threats to the system.

In Rocha Et Al., [10] method, the authors proposed to prevent malicious insiders from compromising passwords & cryptographic keys by VM relocation and obtaining private keys using memory snapshots. None of those methods ensure to achieve holistic security.

In Salvatore Et Al., [11] they proposed a system that confuses the intruders the real information, a totally unique way to deal with securing the cloud using offensive decoy information technology, that they call it as ***Fog Computing***. They used this technology to mitigate malicious insiders, by confusing them with fake worthless data from the actual real user data [2]. They used Support Vector Machine (SVM) for user behavior profiling since SVM is a margin-based classifier, sometimes it leads to misclassification if the value is around the boundary.

In KM Reena Et Al., [12] methodology, they have developed a system with the user profile mapping and decoy technology using encryption technique HMAC with dynamically generated decoy files concept. But the accuracy of detecting of genuine users and performance will be low when considered in multi-cloud environment conditions, large datasets need to be processed.

Considering the above limitations, we posit that the concoction of these features with big data processing framework in intercloud environments improves the performance and accuracy compared to existent.

3. Proposed methodology

In this paper, we propose a new methodology which identifies malicious intruders on monitoring the abnormal access patterns through User Behavioural Profile Mapping with the help of data mining algorithms and map reduce. Upon the detection of intrusion, we invoke the decoy data to the attacker and the owner has been informed of the unauthorized access. Whenever an intruder tries to access the data of owner, our system will detect an abnormal pattern of the data access and consequently creates a decoy file with the same filename by scrambling down the real content of the file to an intruder with bogus information.

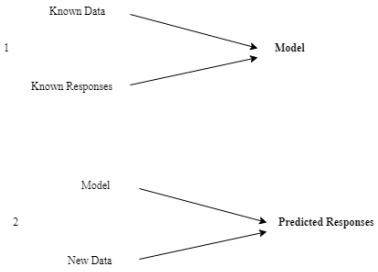
In order to achieve the proposed objective, we divide the whole system into the following modules and each module comprised of distinct algorithms.

User profile mapping

It is always expected that the access of the users' data in the cloud will always exhibit normal means of excess. User Profile Mapping is the method of detecting intruders by comparing with already existing patterns of continuous normal behavior monitoring such as how much a user accesses their data in the cloud. Depicted behavior-based security mostly used by cops in fraud detection. In a large proportion of circumstances, fraud profiles include:

- Multiple attempts to log in the account.
- Huge amounts of information requests within no time.
- How often a doc typically read/write.
- Password & Keys trails.

Machine Learning (ML) algorithms have picked up lots of consideration in recent times, it comes under the field of Artificial Intelligence, which is associated with the study of systems that can learn from data. ML makes a prediction on the data, based on known properties, and learns from the training data. ML algorithms are classified into two types i.e., Supervised and Unsupervised learning. Supervised Learning takes known set of input data and known responses to the data and seeks to build a predictor model that generates reasonable predictions for the response to new data [13].



Hadoop is an open-source framework based on Java programming language used to efficiently store and process the large set of data in a parallel manner over distributed computing environment across the cluster of commodity hardware. It consists of two key components, Hadoop Distributed File System (HDFS) and MapReduce [14]. HDFS is a distributed file system intended for storing very large files with streaming data access patterns, which provides the capabilities to store big data efficiently. MapReduce is a massively parallel processing technique for processing data which distributes a task across multiple nodes.

Naïve Bayes is a supervised learning algorithm widely used for classification high dimensional training datasets.

Our proposed methodology implements the logic of a Machine

Learning algorithm Naïve Bayes using Hadoop's MapReduce function for user behavior mapping on the basis of the following attributes:

1. Login Times
2. Session Times
3. Uploads Count
4. Downloads Count
5. How many files are read and how often.
6. File search or access patterns

Let's define notations which will be using for the rest of the paper.
Users (U_i) = U_1, U_2, U_3, …. U_n
LoginDetails (U_{id}) = U_{1d}, U_{2d}, U_{3d} …. U_{nd}
AnonymousActivites (A_i) = { "invalidPassword", "multiple_LoginAttempts", "wrongLoginTime",.. }
OriginalFiles (O_{ij}) = O_{11}, O_{12}, O_{32}, …. O_{mn}
DecoyFiles (D_{ij}) = D_{11}, D_{22}, D_{32}, …. D_{mn}

Algorithm for detecting user profile mapping

Objective: To detect the behavior of the user Ui

BEHAVIOR 0 -> LEGAL;
BEHAVIOR 1 -> ILLEGAL;
U_i -> CurrentUser
LoginDetails(U_{id}) ->
With all the activities of the User(U_i)
while TRUE do
if AnonymousActivities(A_i) then
++LoginDetails(U_{id})
else
Continue;
end
if LoginDetails(U_{id} < Threshold) then
BEHAVIOR(U_i) = 0;
else
BEHAVIOR (U_i) = 1;
end
end

Decoy technology

We are using Decoy Technology for sending the decoy data to the intruder whenever our system detects the user's behavior as illegal. Decoy data is like garbage data or bogus information which is used to detect unauthorized access information and to confuse the attacker as they assume that this as original data. Thus, users' data is guaranteed to be secure from the intruders by implementing these two features:

- By evaluating whether data access is legitimate or authorized when the strange behavior is detected.
- By creating confusion to the attacker with useless garbage data, which is provided by the decoy files.

Algorithm for Decoy File Generation

Objective: To generate a decoy file
BEHAVIOR 0 -> LEGAL;
BEHAVIOR 1 -> ILLEGAL;
U_i -> CurrentUser
LoginDetails(U_{id}) ->
With all the activities of the User(U_i)
procedure DownloadFile(U_i, file)
if User(ILLEGAL) then
Download ->
DecoyFile(U_i);
else
Download ->
OriginalFile(U_i);
end
end

4. Workflow of the proposed system

Our proposed system involves Hadoop setup at the server, in which cluster of commodity hardware is used to save users data. The data is saved as fragments in different DataNodes, and the NameNode consists of metadata which has the information about where the actual data is saved data nodes. Users' behavior is saved in the database and the detection of abnormal access behavior is done using Naïve Bayes algorithm and MapReduce. Whenever illegitimate access is detected, decoy data is provided and the owner will be notified of this. In case, if the genuine user is being detected as an attacker, OTP verification resolution is proved on the second-time request for the same file. The following Figure 1. shows the proposed system architecture.

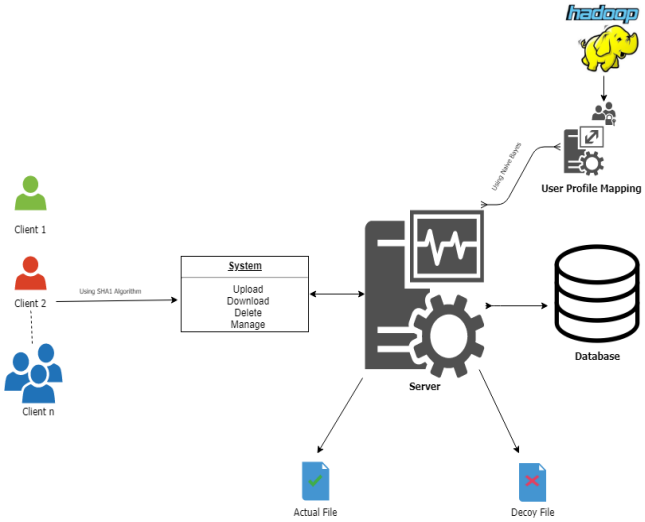


Figure 1: Architecture of the proposed system

The workflow of the proposed architecture is as follows:

- The user creates an account; the registration and authentication are stored in the hash-based format in the database using SHA-1.

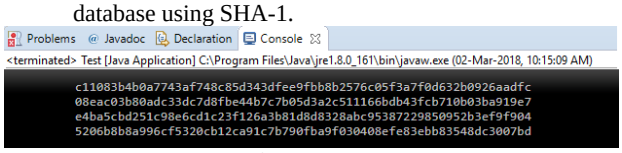


Figure 2: Generation of hash code using SHA-1 Algorithm

- When the user tries to log in, if the hash of the entered password matches with the hash of registered password the user will be redirected to their home page otherwise invalid credentials error notifies.

- For User behavior profiling, parameters like duration, upload & download rates etc., will be considered to detect the user is valid or invalid.
- Whenever the user uploads any file, that is fragmented into chunks using Hadoop framework into the database, on the request for the file, it is fetched from the database using decryption key and sent to the user if he's detected valid otherwise a decoy file will be sent. And the alert will be sent to the owner.

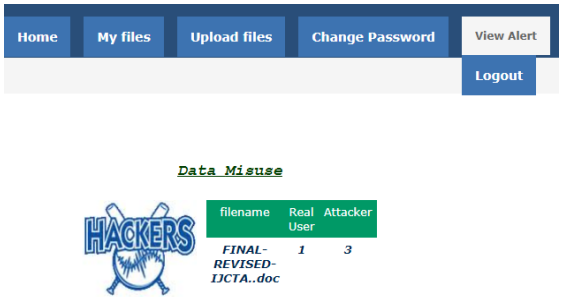


Figure 3: View Alerts to the Owner

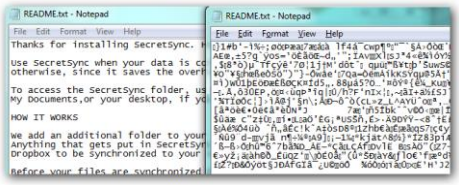


Figure 4: Comparison of Original and Decoy files

5. Applications

- Useful in Governmental Sectors or Organizations in securing citizen or employee databases from intruders who intend to duplicate someone's identity.
- Software Companies can secure their projects that are deployed in Cloud.
- Banking Sectors can make use of it to secure their customer details and transaction details.
- Useful in preventing data theft attacks against the data saved by the normal users in cloud services like DigiLocker etc.

6. Results & analysis

We have analyzed the proposed algorithms of user behavior profile mapping and decoy technology with a distinct type of possible anonymous behaviors on 10 machines to measure how effectively our methodology is able to find the current is the user is genuine or anonymous and on the basis of that, we will be providing the file.
We are drawing the results using a statistical approach. Let's consider X(n) be the number of times we are able to predict correctly whether the user is genuine or anonymous. Let's define another term Y(n) be the number of times our prediction went wrong. Thus, accuracy is being calculated by using the following simple formula:

X(n) -> True Cases

Y(n) -> False Cases

Accuracy = X(n)/(X(n) + Y(n))

We tested our system with 10 students, each of them tried login into our system 40 number of times. The following Table 1. and Figure 5. shows the total number of correct predictions i.e. identification of attacker and non-attacker conditions correctly.

Table 1: Accuracy of the System

S.No.	No. of Correct Predictions	Accuracy
1	38	(38/40) * 100 = 95
2	39	(39/40) * 100 = 97.5
3	35	(35/40) * 100 = 87.5
4	37	(37/40) * 100 = 92.5
5	38	(38/40) * 100 = 95
6	37	(37/40) * 100 = 85
7	39	(39/40) * 100 = 97.5
8	36	(36/40) * 100 = 92.5
9	39	(39/40) * 100 = 97.5
10	38	(38/40) * 100 = 95

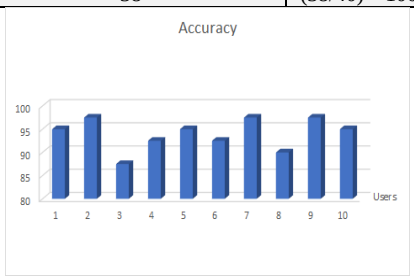


Figure 5: Graphical representation of proposed algorithm's accuracy

As we can notice that, in most of the cases our system was able to detect the behavior of the users correctly. We also compared the results of our algorithm to the previously defined algorithm on the basis of a correct number of predictions of user behavior, with the most number of correct predictions defines the superiority of our algorithm as shown in Table 2. and we can clearly observe Figure 6. our algorithm outsmarts the other one in most of the cases.

Table 2: Comparison of correct prediction values between proposed & previous algorithms

Present Value	Previous Value
95	95.2
97.5	97.5
87.5	87.5
92.5	90
95	85
85	92.5
97.5	90
92.5	82.5
97.5	95
95	92.5

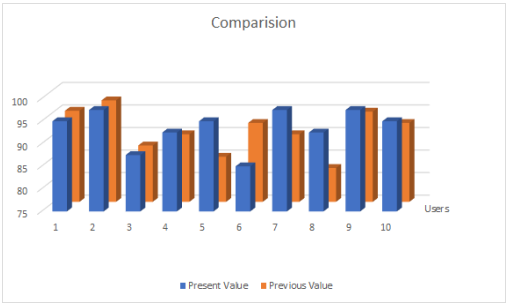


Figure 6: Graphical representation of proposed algorithm's superiority

7. Conclusion

As the increasing data theft attacks have becoming a severe threat to cloud service providers, the proposed approach helps in minimizing data theft over the illegitimate access by monitoring the user behavior and inundate the malicious insider with decoy information. Since our system makes use of both the algorithms SHA1 & Naïve Bayes, which possess higher accuracy and precision rate spreads across two-level security structure compared to the others. This results in Cloud Security at an unprecedented level.

References

[1] Yusop ZM & Abawajy JH, "Analysis of insiders attack mitigation strategies", *Procedia-Social and Behavioral Sciences*, Vol.129, (2014), pp.611-618.

[2] Shiyana S, Bhandare A & Yelwande D, "Fog Computing: Software decoys for insider threat", *SSRG International Journal of Computer Science and Engineering*, (2015).

[3] Botta A, De Donato W, Persico V & Pescapé A, "Integration of cloud computing and internet of things: a survey", *Future Generation Computer Systems*, (2016), pp.684-700.

[4] Cloud Security Alliance, "Top Threats to Cloud Computing Plus: Industry Insights", (2016). https://cloudsecurityalliance.org/group/top-threats/, 2016.

[5] CERT, "Insider Threat". https://www.us-cert.gov/taxonomy/term/58/all.

[6] IBM, "Cyber Security Intelligence Index", (2015). Essextec.com-Cyber-Security-Intelligence- 2015.

[7] The Verge, "Yahoo says all 3 billion user accounts were impacted by 2013 security breach", (2017). https://theverge.com/2017/10/3/16414306/yahoo-security-data-breach-3-billion-verizon, 2017.

[8] Wikipedia, "iCloud leaks of celebrity photos", https://en.wikipedia.org/wiki/ICloud_leaks_of_celebrity_photos.

[9] New York Post, "Michelle Obama's passport leaked by a hacker", (2016). https://nypost.com/2016/09/22/michelle-obamas-passport-published-online-by-hacker

[10] Francisco R & Miguel C, "Lucy in the Sky without Diamonds: Stealing Confidential Data in the Cloud", *1st International Workshop on Dependability of Clouds, Data Centers, and Virtual Computing Environments*, (2011).

[11] Stolfo SJ, Salem MB & Keromytis AD, "Fog computing: Mitigating insider data theft attacks in the cloud", *IEEE Symposium on Security and Privacy Workshops (SPW)*, (2012), pp.125-128.

[12] Reena KM, Yadav SK & Bajaj NK, "Security Implementation in cloud computing Using User Behaviour Profiling and Decoy Technology", *International Conference on Inventive Communication and Computational Technologies*, (2017).

[13] Srujana R & Murty GSN, "Evaluating the Effectiveness of Classification Algorithms Based on CCI", *International Journal of Innovative Research in Science, Engineering, and Technology*.

[14] Apache, "Hadoop Documentation". https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html.