DETECTING INSIDER ATTACKS SEQUENCES IN CLOUD USING FRESHNESS FACTORS RULES

Paramjit Singh[1], Jasmeet Singh Gurm[2]
[1]Research Scholar (M.Tech.), [2]Assistant Professor
Department of Computer Science & Engineering
RIMT- Institute of Engineering and Technology
Mandi Gobindgarh, Punjab, India

ABSTRACT
Sequence mining algorithms can be classified into mainly four ways, viz, apriori-inspired algorithm, pattern matching and pattern growth, pruning and last but not lest the combination any of these. Hence, for each category of these algorithms a suitability of these algorithms for detection of the Insider attack was carried out. These algorithms can be detecting the abnormal patterns in daily routines of the cloud uses with good level of accuracy. As per current status of the state of art in this area, it was found that combination of pattern growth with freshness factors are best suited for identification of insider attack in cloud. As using optimally sized data structure representations of the sequence database necessitates the need of Freshness parameters to be inbuilt in these algorithms. This research work is based on the innovative usage of sequence mining algorithms for detecting Insider attacks in cloud network. As per current work we have found that combinations of pattern growth with freshness factors are best suited for identification of insider attack in cloud. This is apparent from the results from the series of experiments. This implementation basically helped in building an optimally sized data structure representations of the sequence database that necessitates the need of Freshness parameters incorporation in insider attack detection algorithms. It was found that our proposed algorithm is better in terms of memory usage and accuracy in finding the abnormal sequences for detecting the insider attacks.

KEYWORDS - *Inside attack, Sequence mining, Freshness of data.*

I. INTRODUCTION

Intrusion Detection System needs to process large amount of data in one go. As time passes the achieves need higher levels of availability and scalability of processing abilities. Such design choices typically exhibit a tradeoff in which data freshness is sacrificed in favor of reduced access latencies. There are many possible strategies for efficiently handling the recourses so that to strike a fine balance between both the quality of service (QoS) and quality of data (QoD) [1] to be used for analysis in detecting malicious activities in cloud ecosystem. The cloud networks [2] data stream come from time varying processes occurring at different cloud end points , it is not enough to guarantee usefulness of the data for finding malicious activities with mix of old and new data ; we also must ensure each transaction under observation is fresh, so that real time response chain can be build . Informally, data freshness implies that the data is recent, and it ensures that no adversary replayed old messages.

In this paper, we analyze the aspects that have impact in the freshness of datasets [3]for insider attack detecting. Then , the further sections demonstrates a framework that working on Freshness rules for optimization of the Intrusion detecting system [4] performance with respect to its response to malicious activities using sequence mining .Since, the malicious insiders create a bigger threat during a cloud computing surroundings, and the customers don't have a transparent read of supplier policies and procedures. For example, worker access, worker watching, policy compliance and hiring standards/practices area unit usually not clear to customers. Malicious insiders [5] will gain

unauthorized access [6] into organizations and their assets. Some threats embody complete harm, money impact and loss of productivity.

The rest of the document is organized as follows: Section 2 analyzes the aspects involved in freshness evaluation and presents basis for understanding the need for incorporation such factors in intrusion detection system . Section 3 describes the framework, which is used in section 4 to evaluate detection process in a particular scenario. Finally, section 5 concludes with our general remarks.

Section 2: Aspects that influence the Analysis of the "Insider Attack" [7]: This section discusses the factors that must be fine-tuned to get highly optimized Intrusion detection system.

Table no 1.1: Variables impacting Insider Attacks

S.No	Metric	Description	Influence on Detection
1	Currency	Time since data was extracted from the sources, basically time difference between the query time and extraction time	More the extraction time and Query, more is the delay in Alert and Response, which means IDS is slow in detection.
2	Obsolescence [8]	The number of updates since the data was extracted since execution time	In cloud, this is every critical, if the real time activities transaction is too high; there is difference in real time analysis. The IDS may miss some data
3	Freshness Rate [3]	The percentage of tuples up to date (have not been updated since extraction time)	If the percentage of freshness less, the database /sequences will undergo large scan time. There by slowing the alert and response time
4	Timeliness	The difference between the last update time since extraction time	It is desired that this must be minimal for best results
5	Size of Data Set	Total number of tuples /rows/items to be scanned for finding the insider attacker in database	Larger datasets means more time in scanning them and slower responses.
6	Size of Sequence	Encoded tuple for identification of a transaction occurrence in cloud	Larger sequences need more time for processing.
7	Mining Time [9]	Time taken by an algorithm to find particular set of sequences	It is desired that this must be minimal for best results

II. LITERATURE SURVEY

After systematic review using high impact research papers on "Intrusion detection systems" and "Insider attack" following parameters were found [Table no 1. 2] to be critical for building a state of art detection system that detects malicious cloud insiders [10].

Table no. 1.2: Parameters for Detecting Insider Attack

S.no	Parameter in context of sequence mining	Description	Requirement for Insider Attack Detection Algorithm
1	Sequence [11]	An ordered collection of Item sets or Set Predicates is defines as Sequence.	The sequence essentially must have all possible codes of parameters from which insider attack can be detected
2	Size of Dataset	Number of Items sets /Row/Tuples.	These will be equal to number of rows 'r' multiple by columns 'c' or r*c at any given time, where c is the influencing parameters.
3	Type of Dataset & Encoding of sequence events	Encoded set items like a,b or 1,2,3 or combination a1,b1 which means dataset will consist of alphanumeric types .	Basically it is "String mining" In which a limited alphabet for items that appear in a

			sequence, but the sequence itself may be typically very long. This is typical characteristic of "insider attack" sequence.
4	Memory Requirement	Memory usage (in megabytes) with respect to Threshold	It is be optimized and remain low for implementation of the algorithm.
5	Scalability Requirements	This requirement basically try to find the influence of the number of transactions on execution	The detection system must be scalable to deal with huge amount of data as well must be able to allocate MIPS on the fly to scale up
6	Execution Time	It is the time taken for IDS program to get execution.	It is should be minimum for an alert system to be successful in informing the system administrator about the insider attack
7	Running Time	In this requirement assessment the varied threshold is to consider the influence on execution time. The number of time periods is also considered (5, 25 and 50 periods).	Running time must be minimized and the impact of longer running time must be observed and marked, so that typical synchronized alert system can be build.
8	Database Scans	Full database , Partial ,Indexed scans , Multiple Scans	Incremental and Interactive mining is basic requirement for "insider attack" sequence and it volume of data.
9	Size of Sequence	The sequence length defines the amount of dataset attributes or size of transaction or simply the total events that can be considered as transaction for analysis.	This is directly related to the number of parameters under observations for detecting the attack
10	Number of Rules , Sequence Rule	A Sequence Rule [12] model gives regulations for various sets or items. It basically describes the relationship between two sequences.	There should be optimal number of rules, so that computation over head is minimal.
11	Number Of Transactions	Frequency of activities considered as a group.	The number of objects based on which the group was build on.
12	Number Of Items Per Transaction	It defined as maximum number of activity (e.g., visits of particular cloud service) per object/ cloud subject/user	If there are abnormal number visits to a particular cloud service by a user or both , it is an indication of some attack
13	Number Of Transaction Groups	When the visits or a metric are grouped in some logical manner.	If a particular group have number of abnormal visit by cloud or both ., it may be indication of some adversity
14	Confidence	It the ratio of the number of objects in the data for which the antecedent and consequent Sequences hold true, to the total number of objects in the data.	How much proportion of data elements are confident on your hypothesis.
15	Support	Probability of the consequent following the antecedent. Calculated as the number of occurrences of a sequence rule divided by the number of occurrences of the antecedent.	Probability value bears all or part of the weight of; hold up for the hypothesis in question. Whether, it is an attack or not.

Main Types of Inside Attack Detection Algorithms [13]:
First things first, there is need to select the model based on which we will be detecting the adversity, for this model selection task needs to be conducted from the sets of candidate models, given data. This involves the design of series of experiments such that the data collected is well-suited to the problem of model selection. Given the choice of models evaluation with experimental results the candidate models having similar predictive or explanatory powers are considered the best choice for detecting the adversity. Following models may be considered, which may fall in category of exploratory in nature or a scientific method of inquiry

Table No. 1.3: Description of Detection Methods

S.NO	Method of Detection	Description
1	Statistical Analysis Method [14]	It is a way in which data or samples can be explained/ described &précised to draw conclusions for detecting an adversity for example
2	Data Mining Method [15]	It is a way of finding knowledge by sending queries to get useful information to check insider attack for example.
3	Data Stream Analysis [16]	It refers to the analysis of data , which is coming in real time , being produced by an ongoing process like "cloud services" running to compute weather reports or stock exchange reports
4	Machine Learning Method [16]	In case machine learning we are looking for patterns in data for learning, classification or grouping to reach at some decision .For example to learn inside attacks activity behavior
5	Probability Based Method [17]	These methods are basically based on the probability to find probable conditions with some level of confidence e.g. probability of insider attack, if some pattern of activities is found for a particular cloud user .
6	Sequence Mining	It can help to get both infrequent or frequent sequence patterns of events which can lead to conclusive proof that there is an attack going on in the network
7	Ranking {Trust, Reputation Voting } [18]	In this method, points are given to the objects that gain trust over time either by using reputation algorithm or voting.
8	Thresh holding [19]	In this method a dynamic values with heuristics evaluation may be used to check insider attackers

III. RESEARCH GAPS

After the systematic survey, it was found that a sequence mining algorithms can be used to fine tune with freshness parameters and factors to avoid full database scans and increase the alert and response time for detecting malicious activities in cloud. It was also found that some sequence mining algorithms like prefix already in use for detecting malicious activities in the Intrusion detection systems. However, these algorithms can be increase their accuracy and responsiveness in incorporating Obsolescence metrics, Freshness factors. Limited works have been reported in this context of Insider attack detection using such factor rules or metrics.

Scope of work:

- **i.** Develop a simulated cloud environment with normal and insider attack scenarios.
- **ii.** Using Freshness Rule Based data mining technique for detection of Insider attack in cloud.
- **iii.** Evaluate performance with previous work

IV. IMPLEMENTATION

In this section, we detail the steps of implementation in conducting the research work. Basic flow the detection process is as follows.

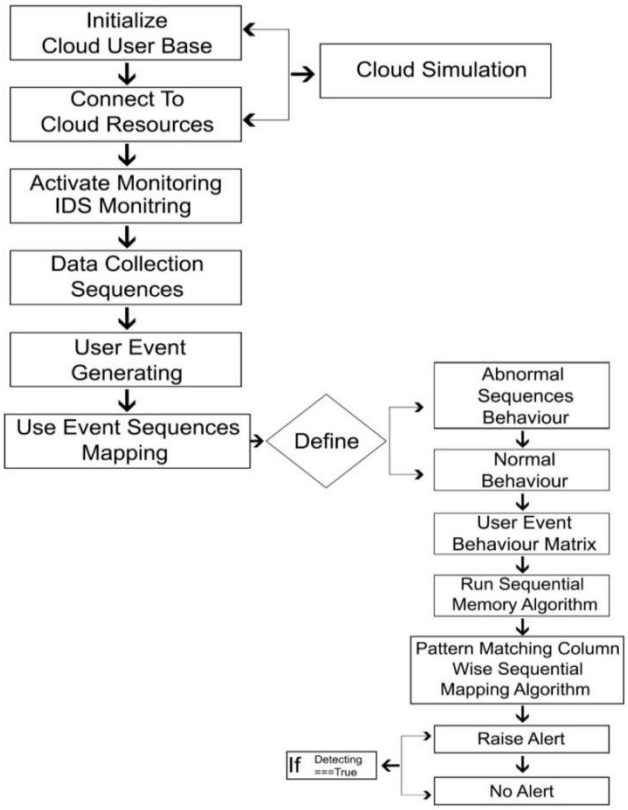

Figure 1.1: Block Diagram of IDS

Step 1: This step includes the building of the cloud simulated ecosystem, for this we have used discrete event model of simulation .The simulation was built in core java 8 and main high lights of the cloud setup include following along the facility to imitate malicious user activities /transactions and sequences.

Creation of Datacenter: A data center is the operational center having physical machines. These machines have Virtual machines and all the other facilities to process work given by broker.

Cloud User/Broker Entity: It is entity which accumulates work for the data center(s). It may be part of group computers which are sending work to the data center.

VM Allocation and Migration Policies Units: This term basically means how the work from the broker is distributed to virtual machines of the data center.

Step 2: Implementation of Intrusion Detection System: Collection and Aggregation of Sequences: The activities of the users are maintained in central database updated from various cloud end points at regular intervals. The activities are encoded in numerical encodes and ranked by the sequence mining algorithm , which scan the database for most infrequent patterns i.e abnormal using Freshness Factors rules , which is basically calculated based on sliding window time series and possible response time statistics.

V. ANALYSIS OF COLLECTED DATA

Freshness Factor based Sequence Mining algorithm is used to detecting the abnormal activism of cloud users. Each transaction may consist of 2 or more activities. The sequence may be as small as 2 codes or may be as large as 6-7 activities or even more but for a threshold value is recommended to keep the transaction size intact. Each transaction/Event Sequence is a set of items (symbols) .The main steps in algorithms are as follows

Proposed Algorithm (Freshness Factors Algorithm):

a) Key Advantages: The main advantage of our algorithm is that is checks all the values , that are recent or simply fresh and avoid scanning old data again and again .This way the algorithm runs fast and uses less memory .

b) Input for Checking the Insider attack detection: The main input is sequence of the cloud user activities , At least 2 activities can be defined a cloud user activity for analysis However , threshold limit is consisted , so that transaction size does not exceed much and make the data base huge and slow for scanning . Each transaction/Event Sequence is a set of items (symbols) as shown in Figure1.1

Step 1: This step consists of appending the latest transaction T_k to the current transaction list or data base CTL.

Step 2: In this step new count are added and updated. For each "item set" or "sequence set" that appears in the new transaction T_k with an entry (e, f, t), if, there is a corresponding node is in the observation space lattice, the count f of the corresponding node point is increased by value of one(e.f = e.f + 1) . Then, for ("cloud user activity sequence set") or the new item set "e" induced by the items of the new transaction T_k is inserted into the observation space lattice with an entry (e, f = 1, t = k).

Step 3: Getting a Transaction from the data base scan: In this step the oldest transaction is extracted and entry is made in observation space lattice i.e. CTL, But, if its corresponding node with an entry (e, f, t) is in the observation space lattice, the count f of the corresponding node is updated.

Step 4: This step is for pruning of user activity sequence /item sets for optimization purposes. For each sequence set "e: with an entry (e, f, t) in the observation space lattice , if it has maximum value of support "C" max, then it is pruned .

Step 5: Now , we get the most infrequent or rare sequence of activities , because the insider attacker activities will be less in proportion as compared to the normal user sequences, Here, simply, the item set e with an entry (e, f, t) in the observation space lattice, if its minimum possible support C max are considered.

The Pseudo Logic:

Let f be the freshness rule factor,
Let ds be the size of the dataset.
Let sn be the length of the sequence.
For Each "Sequence" of User Activities in Database
Slice dataset based on freshness rule factor 'f'
For each Sequence tree,
remove the sequence having size larger than "Threshold"
Find most infrequence sequence tree, "CTL"
Add_to_List _Of_InsideAttackers("t");
End

Outcomes of algorithm: The propose algorithm picks up the most infrequent Sequences, by building a first a tree of the sequences and then by removing the duplicate sequences to get list of unique sequences and rank them according to their frequency of occurrences. Most algorithms try to find most frequent items, however our need to find most infrequent.

The Output Format:

The output of the algorithm is as follows

2 3 5 #SCORE: 0.0000019860383547
2 2 2 #SCORE: 0.0300002131312072
2 1 2 3 #SCORE: 0.033335643690463074

VI. ALERT AND RESPONSE SERVICES

This is when, some activity sequence transaction is found be infrequent. A alert message is send and performance of the algorithm is measured in term of time repose and accuracy of detection of insider attack

VII. RESULTS

In this section, we shall discuss a testable hypothesis (that malicious activities of a cloud user can be traced by analysis of his sequence of activities) that can be answered experimentally by doing simulation of in cloud environment, this was only possible by collecting samples of the cloud users events and then by running proposed algorithm we try to detect if there is some abnormal sequence of events. The accumulated observations about this are shown graphically in further sections of this chapter. These observations are then analyzed to yield an answer to the key question regarding to time response, memory and accuracy of the results

Table no. 1.4: Evaluation Parameters

S.No.	Factor	What it will Evaluate?
1	Total Number of Transactions in Cloud	This helps to determine the size of dataset for conducting analysis , it reflects how much data algorithm can handle in reason repose time to report adversity
2	Total Number of Sequences that are Abnormal (Malicious)	It is hard to find "abnormal "sequences /transaction when there is huge proportions of normal activities as compared to normal.
3	Total Number of Sequences that are Normal :	This means what proportion of cloud activities are routine and normal.
4	Total Time in Identification of Malicious Patterns of Activities	Time in detection of malicious activity. It is total time in mining the infrequent or malicious patterns.
5	Total memory Consumption in processing for identifications of Malicious Pattern Activities	It measure of how much memory get used up when algorithms runs to find abnormal activity .It is always desired that minimum ram must be used .
6	Execution Time	Total taken for the intrusion detection system to run or scan or do analysis of 1 batch of transactions,. The size of batch may consist of 10,000 sequence tuples.
7	Composite Freshness Factor	It is a derived metrics which measures the size of dataset and remove configurable number of tuples from the current dataset, as it has become Obsolescence or ready for archiving.

Graphical Presentation of the Results

a) Mean of Total Sequences or Instances: This graph describes the finding a unusual or infrequent malicious activities that can be termed a "insider attack". By using this metric, we are able to investigate the slowness of the database. But, we the use of our freshness algorithm, same dataset will not need multiple scan as only fresh values are scanned. The graph below show that our algorithm scans more items in less time due to the fact the sequence tree is smaller due to freshness factor.

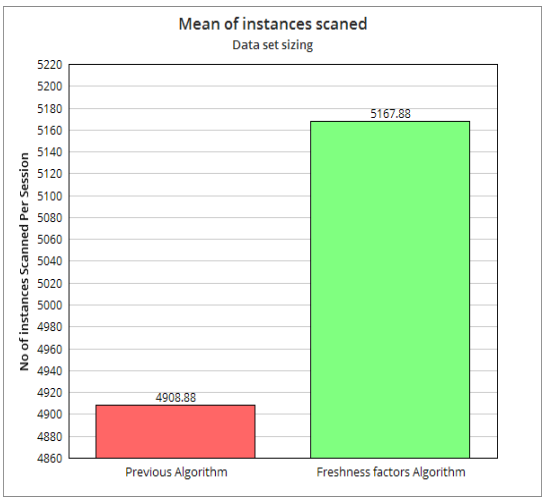

Figure 1.2: Mean of Total Sequences or Instances

b) Mean of Sequences or Instances found to be normal: This graph show average number of sequences that were found to be normal routine .This metric will help to find, whether the algorithm is able to differentiate between normal and abnormal sequence.

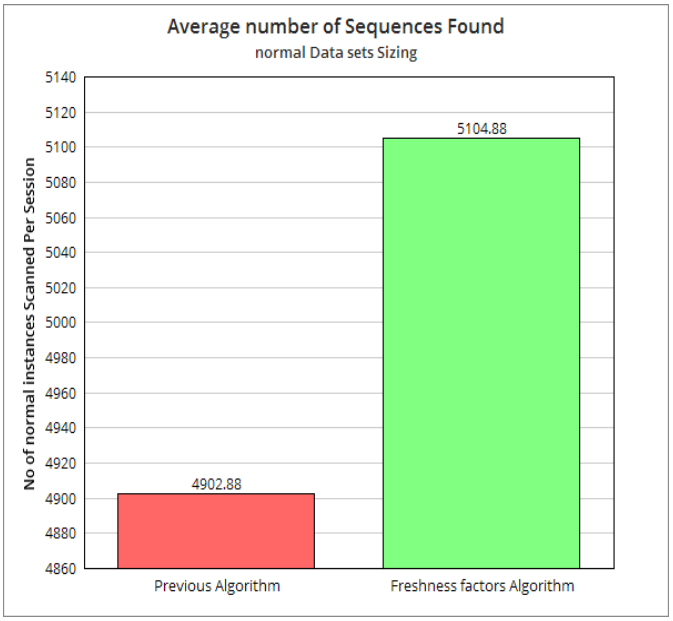

Figure 1.3: Mean of Sequences or Instances found to be normal

c) Mean Abnormal Sequences Number: This graph shows the results from the both algorithms for finding most infrequent instances of sequences that normally do not happen or are malicious in nature. It shows that our proposed algorithm is able to find more number of sequences that are abnormal.

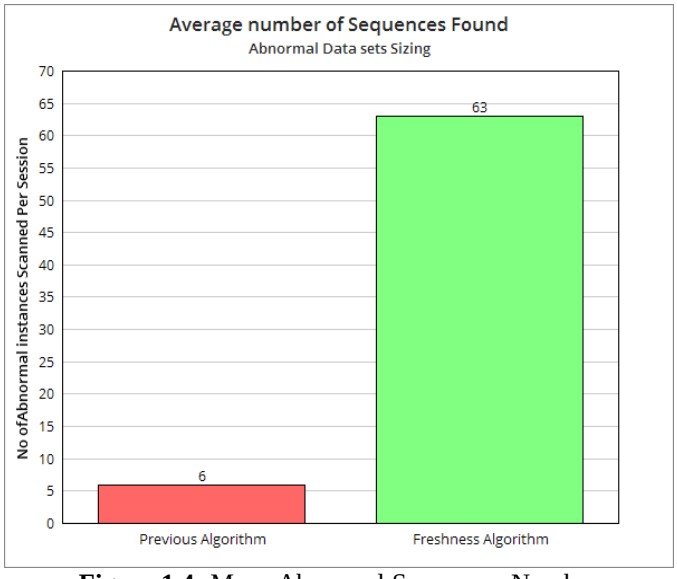

Figure 1.4: Mean Abnormal Sequences Number

d) Mean Time in Inside Attack Detection: This following graph shows how much it takes to find the both the "normal" and abnormal sequences. It is clear from the graph below that the proposed algorithms takes on less time in finding the abnormal activities of the cloud user .

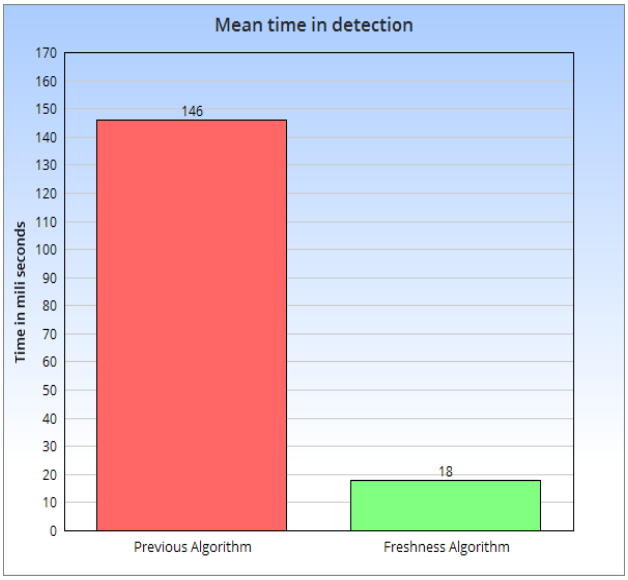

Figure 1.5: Mean Time in Inside Attack Detection

Interpretation of all Outcomes and graphs:

a) It can also be observed that the average time in finding the malicious activity is very small making it highly efficient.

b) It can also be observed that the mean memory consumption is also very low making the data processing memory efficient.

c) It can also be observed that the mean memory consumption is also very low making the alert response time less.

d) It can also be observed that the mean time in finding the malicious activity is very less making "response time for adversity defense fast".

e) It is also clear from out working the if access policy is not effective, the algorithm might find malicious patterns, it would remain a pain in the neck.

f) From graph, we can interpret that there are large number of cloud users sequences (5000-6000), defining the behavior of cloud users. This amounts to a huge database scan in one go.

g) Since, the number of cloud user transactions are around 5000-6000 per session, the ratio of normal and abnormal is more towards normal actually, it makes the work of algorithm challenging.

VIII. DISCUSSION AND CONCLUSION

In this research effort, we have been able successful to develop an algorithm that catches abnormal activities of the cloud user, the implemented algorithm requires less memory space for storing and searching. The user activity give rise to many types of sequences These sequences are based on the events that are generated while the user is interacting with the cloud services, the algorithm automatically skips or slides to new data set for fresh scan based on the rules that calculate the 'freshness' of the user event sequence. However, no matter how strong the algorithm may be, if human nature is taken in to account, human are more prone than the software sometimes, therefore, We need to make sure there are checks and balances in place and that sensitive information is accessible only to those who truly need it in order to be able to do their job properly and, more so, we need to make sure it's easy to revoke access to sensitive information at a moment's notice, especially on mobile devices. Google offers the ability to remotely wipe mobile devices, and more companies are following suit. Check into these solutions sooner rather than later. The outcome shows that it is conceivable to identify insiders that disguise in the framework by watching their conduct designs.

In summary, we can say, this research covered the problems related to finding right kind of parameters and freshness factor algorithm that would be suitable to detection of insider attack. The process consists of identification of the events point that needs to be traced for benchmarking "normal event" or abnormal event. This step can also be referred as data pre-processing phase , a step in which

event sequence code are selected , defined and cleaned to form final transformed sequence structure to analysis . Once, this step is done, the cloud user activity data are now ready for pattern discovery and finding rare abnormal events. This is where we apply algorithms to identify knowledge embedded in data, and to evaluate the discovered knowledge, Here, the Knowledge is "Presence of Malicious Cloud User". Last, but not least, step is to find explanations, constructs and evaluation of the discovered knowledge (Presence or absence of Malicious User), and to response in such a manner that there is no false alarm.

IX. FUTURE SCOPE

These days, there are more advance insider attacks that have come in to play. These attacks also happen due to insider's malicious coordination with outsiders. Therefore for future scope we suggest that this work may be extended to detect new type attack called social engineering attacks. For future direction we also suggest that sequence mining can become more fast in terms of retrieving data if we use databases that support object serialization and single phase query support like Ceph database

REFERENCES

[1] G. F. Anderson, D. A. Selby, and M. Ramsey, "Insider attack and real-time data mining of user behavior,"
[2] Chirag Modi et al., "A survey of intrusion detection techniques in Cloud," Journal of Network and Computer Applications, vol. 36, no. 1, pp. 42-57, jan 2013.
[3] Ari Juels and Alina Oprea, "New approaches to security and availability for cloud data," Communications of the
[4] Poonam Sinai Kenkre, Anusha Pai, and Louella Colaco, "Real Time Intrusion Detection and Prevention System," in Advances in Intelligent Systems and Computing.: Springer Science Business Media, 2015, pp. 405-411.
[5] Syam Kumar Pasupuleti, Subramanian Ramalingam, and Rajkumar Buyya, "An efficient and secure privacy-preserving approach for outsourced data of resource constrained mobile devices in cloud computing," Journal of Network and Computer Applications, vol. 64, pp. 12-22, apr 2016.
[6] Qussai Yaseen, Qutaibah Althebyan, Brajendra Panda, and Yaser Jararweh, "Mitigating insider threat in cloud relational databases," Security and Communication Networks, vol. 9, no. 10, pp. 1132-1145, jan 2016
[7] T. Gunasekhar, K. Thirupathi Rao, and M. Trinath Basu, "Understanding insider attack problem and scope in cloud," in 2015 International Conference on Circuits, Power and Computing Technologies
[8] Rongxing Lu, Privacy-Enhancing Aggregation Techniques for Smart Grid Communications.: Springer International Publishing, 2016.
[9] Chunyang Yu, Wei Zhang, Xun Xu, Yangjian Ji, and Shiqiang Yu, "Data mining based multi-level aggregate service planning for cloud manufacturing," J Intell Manuf, dec 2015
[10] Mahmoud Barhamgi, Arosha K. Bandara, Yijun Yu, Khalid Belhajjame, and Bashar Nuseibeh, "Protecting Privacy in the Cloud: Current Practices, Future Directions," Computer, vol. 49, no. 2, pp. 68-72, feb 2016.
[11] Sanchika Gupta and Padam Kumar, "An Immediate System Call Sequence Based Approach for Detecting Malicious Program Executions in Cloud Environment," Wireless Pers Commun, vol. 81, no. 1, pp. 405-425, oct 2014.
[12] Nikolaos Pitropakis, Aggelos Pikrakis, and Costas Lambrinoudakis, "Behaviour reflects personality: detecting co-residence attacks on Xen-based cloud environments," International Journal of Information Security, vol. 14, no. 4, pp. 299-305, aug 2014
[13] Roberto Pagliari et al., "Insider attack detection using weak indicators over network flow data," in
[14] Mirco Marchetti, Fabio Pierazzi, Michele Colajanni, and Alessandro Guido, "Analysis of high volumes of network traffic for Advanced Persistent Threat detection," Computer Networks, jun 2016.
[15] J. Amudhavel et al., "A Survey on Intrusion Detection System: State of the Art Review," Indian Journal of Science and Technology, vol. 9, no. 11, mar 2016.
[16] Michael Mayhew, Michael Atighetchi, Aaron Adler, and Rachel Greenstadt, "Use of machine learning in big data analytics for insider threat detection," in
[17] Ashish Singh and Kakali Chatterjee, "A secure multi-tier authentication scheme in cloud computing environment," in 2015 International Conference on Circuits, Power and Computing Technologies [

[18] Shanhe Yi, Zhengrui Qin, and Qun Li, "Security and Privacy Issues of Fog Computing: A Survey," in Wireless Algorithms, Systems, and Applications.: Springer Science $\mathplus$ Business Media, 2015, pp. 685-695.
[19] Praveen Kumar Rajendran, "Hybrid Intrusion Detection Algorithm for Private Cloud," Indian Journal of Science and Technology, vol. 8, no. 35, dec 2015.

BIOGRAPHY

Paramjit Singh is currently doing his M-Tech in Computer Science Engineering from RIMT-Institute of Engineering and Technology, Mandi Gobindgarh under Punjab Technical University, Jalandhar. His research interests are in the areas of Cloud Computing, Networking, Data Mining, and Data Warehousing.

Jasmeet Singh Gurm is currently doing his PhD in Computer Science from Punjab Technical University, jalandhar. He is currently a Professor in RIMT-Institute of Engineering and Technology, Mandi Gobindgarh, India. His research interests are in the areas of Networking, High Speed Computing, Data Warehousing, and Data Mining. He has published four computing books, and numerous research articles. He is also a member of I.S.T.E (Indian Society for Technical Education) and other various organizations.