International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Survey: Anomaly Detection in Cloud Based Networks and Security Measures in Cloud Date Storage Applications

Dr. Chinthagunta Mukundha

Associate Professor, Department of IT, SNIST, Hyderabad-501301, India

Abstract: ***Now a days in the cloud networks huge amounts of data are stored and transferred from one location to another location. The data that is transferred from one place to another is exposed to attacks. Although various techniques or applications are available to protect data, loopholes exist. Anomaly detection uses some data mining techniques to detect the surprising behavior hidden within data increasing the chances of being intruded or attacked. Various hybrid approaches have also been made in order to detect known and unknown attacks more accurately. Cloud computing has become one of the most projecting words in the IT world due to its design for providing computing service as a utility. The typical use of cloud computing as a resource has changed the scenery of computing. Due to the increased flexibility, better reliability, great scalability and decreased costs have captivated businesses and individuals alike because of the pay-per-use form of the cloud environment. Cloud computing is a completely internet dependent technology where client data are stored and maintained in the data center of a cloud providers. The Anomaly Detection System is one of the Intrusion Detection techniques. It's an area in the cloud environment that is been developed in the detection of unusual activities in the cloud networks. Although, there are a variety of Intrusion Detection techniques available in the cloud environment, this survey paper exposes and focuses on different IDS in cloud networks through different categorizations and conducts comparative study on the security measures of Drop box, Google Drive and iCloud, to illuminate their strength and weakness in terms of security.***

Keywords: Intrusion Detection, Scalability, Anomaly Detection, Cloud Computing, Security

1. Introduction

Cloud computing is not giving any assurance on data security but it is a requirement in the IT world. The benefits of cloud computing have no in finite end as to what can't be done using the cloud environment due to a variety of deployment model such as Software as a Service, Platform as a Service, and Infrastructure as a Service. The cloud computing technology allows the clients for much more reliable and efficient computing by centralized storage, memory, processing and bandwidth. This allows the cloud uses flexibility in accessing the cloud data over the cloud network. Analyzing the traffic in cloud networks is one of the most important tasks in cloud management to give the assurance of quality of services, Testing the performance of new applications, build accurate network models and detect anomalies in the cloud. The flow of network that is been created by cloud computing systems shows users' behavior in service operation or use. Traffic analysis and the recognition of all significant application flows are important tools for modeling service usage, building up patterns for identifying normal system operations. The cloud computing environment has accrossed numbers of security challenges. Most of them have been solved up to an extent, other security aspects spring up and it's vital to know before organizations switch fully. Intrusion detection system in cloud networks plays a very important role as the active security defense against intruders. Intrusion Detection System (IDS) needs to be employed properly in the cloud networks, because it requires scalability, efficiency and virtualized-based approach in implementation. The users of cloud computing have a limited control over its data and resources that have been hosted on a cloud service provider remote servers. Due to this proposed theory, it automatically becomes the responsibility of the cloud service provider to oversee the IDS in the cloud environment. Additionally, network communication between cloud providers And its customers affects significantly the performance of most cloud-based applications. Analyzing the flow of network traffic provides insights on how applications behave and also their performance in cloud environment. Therefore, it is necessary to develop network traffic measurement and analysis techniques to improve availability, performance and security in cloud computing environments.

2. Anomaly Detection

Anomaly detection is the process of finding the patterns in a dataset whose behavior is abnormal on expected. These unexpected behaviors are also termed as anomalies or outliers. The anomalies cannot always be categorized as an attack but it can be a surprising behavior which is previously not known. It may or may not be harmful. The anomaly detection provides very significant and critical information in various applications, for example Credit card thefts or identity thefts]. When data has to be analyzed in order to find relationship or to predict known or unknown data mining techniques are used. These include clustering, classification and machine based learning techniques. Hybrid approaches are also being created in order to attain higher level of accuracy on detecting anomalies. In this approach the authors try to combine existing data mining algorithms to derive better results. Thus detecting the abnormal or unexpected behavior or anomalies will yield to study and categorize it into new type of attacks or any particular type of intrusions. This survey attempts to provide a better understanding among the various types of data mining approaches towards anomaly detection.

3. Process of Anomaly Detection

There are different anomaly detection approaches exists, as shown in figure 1 parameter wise train a model prior to detection.

Parameterization: Pre processing data into a pre-established formats such that it is acceptable or in accordance with the targeted systems behavior.

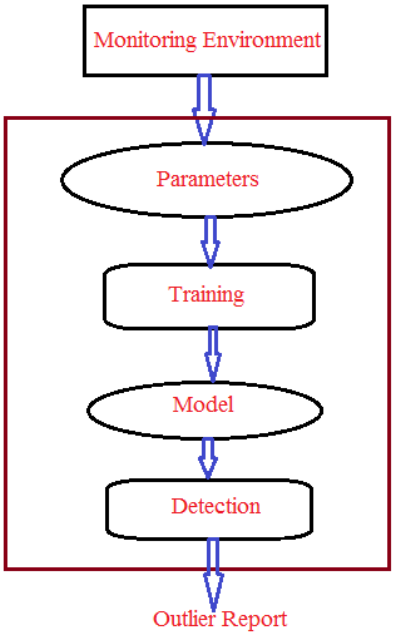

Figure 1: Process of Anomaly Detection

Training stage: A model is built on the basis of normal behavior of the system. There are different ways that can be opted depending on the type of anomaly detection considered. It can be both manual and automatic.

Detection stage: When the model for the system is available, it is compared with the observed traffic. If the deviation found exceeds from a pre defined threshold then an alarm will be triggered.

Managing and analyzing network traffic of large scale cloud systems is a challenging task. The techniques used to monitor and analyze traffic in conventional distributed systems differ from cloud computing systems. In conventional approaches, assumptions are made that network flows follow some patterns, which is acceptable for corporate applications, but cloud applications may have significant changes in traffic patterns.

4. Theoretical Basis and Literature Review

Classification based anomaly detection

Classification can be defined as a problem of identifying the category of new instances on the basis of a training set of data containing observations (or instances or tuples) whose category membership is known. The category can be termed as class label. Various instances can belong to one or many of the class labels. In machine learning, classification is considered as an instance of supervised learning for example learning where a training set of correctly-identified observations is available. An algorithm that implements classification is known as a classifier. It is constructed to predict categorical labels or class label attribute. In case of anomaly detection it will classify the data generally into two categories namely normal or abnormal. Following are common machine learning technologies in anomaly detection:

- **Classification Tree:** In machine learning classification tree is also called as a prediction model or decision tree. It is a tree pattern graph which is similar to flow chart structure; the internal nodes are a test property, each branch represents test result, and final nodes or leaves represent the class to which any object belongs. The most fundamental and common algorithm used for classification tree is ID3 and C4.5 There are two methods for tree construction, topdown tree construction and bottom-up pruning. ID3 and C4.5 belong to top-down tree construction . Further classification tree approaches when compared to naive bayes classification, the result obtained from decision trees was found to be more accurate.
- **Fuzzy Logic:** It is derived from fuzzy set theory which deals with reasoning that is approximate rather than precisely deduced from classical predicate logic. The application side of fuzzy set theory deals with well thought out real world expert values for a complex problem. In this approach the data is classified on the basis of various statistical metrics. These portions of data are applied with fuzzy logic rules to classify them as normal or malicious. There are various other fuzzy data mining techniques to extract patterns that represent normal behaviour for intrusion detection that describe a variety of modifications in the existing data mining algorithms in order to increase the efficiency and accuracy.
- **Naive bayes network:** There are many cases where the statistical dependencies or the causal relationships between system variables exist. It can be difficult to precisely express the probabilistic relationships among these variables. In other words, the former knowledge about the system is simply that some variable might be influenced by others. To take advantage of this structural relationship between the random variables of a problem, a probabilistic graph model called Naïve Baysian Networks (NB) can be used. This model provides answer to the questions like if few observed events are given then what is the probability of a particular kind of attack. It can be done by using formula for conditional probability. The structure of a NB is typically represented by a Directed Acyclic Graph (DAG) where each node represents one of system variables and each link encodes the influence of one node upon another. When decision tree and baysian techniques are compared, though the accuracy of decision tree is far better but computational time of baysian network is low. Hence, when the data set is very large it will be efficient to use NB models.
- **Genetic Algorithm:** It was introduced in the field of computational biology. These algorithms belong to the larger class of Evolutionary Algorithms (EA). They generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, selection, mutation and crossover. Since then,

they have been applied in various fields with very promising results. In intrusion detection, the Genetic Algorithm (GA) is applied to derive a set of classification rules from the network audit data. The support-confidence framework is utilized as a fitness function to judge the quality of each rule. Significant properties of GA are its robustness against noise and selflearningcapabilities. The advantages of GA techniques reported in case of anomaly detection are high attack detection rate and lower false-positive rate.

- **Support Vector Machine:** These are a set of related supervised learning methods used for classification and regression. Support Vector Machine (SVM) is widely applied to the field of pattern recognition. It is also used for an intrusion detection system. The one class SVM is based on one set of examples belonging to a particular class and no negative examples rather than using positive and negative example. When compared to neural networks in KDD cup data set, it was found out that SVM out performed NN in terms of false alarm rate and accuracy in most kind of attacks.
- *Neural Networks:* It is a set of interconnected nodes designed to imitate the functioning of the human brain. Each node has a weighted connection to several other nodes in neighboring layers. Individual nodes take the input received from connected nodes and use the weights together with a simple function to compute output values. Neural networks can be constructed for supervised or unsupervised learning. The user specifies the number of hidden layers as well as the number of nodes within a specific hidden layer. Depending on the application, the output layer of the neural network may contain one or several nodes. The Multilayer Perceptions (MLP) neural networks have been very successful in a variety of applications and producing more accurate results than other existing computational learning models. They are capable of approximating to random accuracy, any continuous function as long as they contain enough hidden units. This means that such models can form any classification decision boundary in feature space and thus act as non-linear discriminate function.

Clustering based Anomaly Detection techniques

Clustering can be defined as a division of data into group of similar objects. Each group, or cluster, consists of objects that are similar to one another and dissimilar to objects in other groups. Clustering algorithms are able to detect intrusions without prior knowledge. There are various methods to perform clustering that can be applied for the anomaly detection. Following is the description of some of the proposed approaches

- **EM Clustering:** This algorithm can be viewed as an extension of k means which assigns an object to the cluster to which it is similar, based on the mean of cluster. In this approach instead of assigning object in the dedicated cluster, assign the object to a cluster according to a weight representing the probability of membership. In other words there are no strict boundaries in between the clusters. Here new mean is computed on the basis of weight measures. When compared to k means and k medoids, EM outperformed them and resulted in higher accuracy.
- **K-Medoids:** This algorithm is very similar to the k-Means algorithm. It differs mainly in its representation of the different clusters. Here each cluster is represented by the most centric object in the cluster, rather than by the implicit mean that may not belong to the cluster. The k-medoids method is more robust than the k-means algorithm in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean. This method detects network anomalies which contains unknown intrusion. It has been compared with various other clustering algorithms and have been find out that when it comes to accuracy, it produces much better results than k-Means.
- **K-Means**: k-Means clustering is a cluster analysis method where we define k disjoint clusters on the basis of the feature value of the objects to be grouped. Here, k is the user defined parameter [9]. There has been a Network Data Mining (NDM) approach which deploys the K-mean clustering algorithm in order to separate time intervals with normal and anomalous traffic in the training dataset. The resulting cluster centroids are then used for fast anomaly detection in monitoring of new data.
- **Outlier Detection Algorithms:** Outlier detection is a technique to find patterns in data that do not conform to expected behavior. Since an outlier can be defined as a data point which is very different from the rest of the data, based on certain measures. There are several outlier detection schemes. User can select any one of them on the basis of its efficiency and how he can deal the problem of anomaly detection. One of the approach is Distance based Approach. It is based on the Nearest Neighbour algorithm and implements a well-defined distance metric to detect outliers. Greater the distance of the object to its neighbour, the more likely it is to be an outlier. It is an efficient approach in detecting probing attacks an Denial of Service (DoS) attacks. Other one is Density based local outlier approach. Distance based outlier detection depend on the overall or global distribution of the given set of data points. The data is not uniformly distributed thus the distance based approach encounter various difficulties during analysis of data. The main idea of this density based method is to assign to each data example a degree of being outlier, which is called the Local Outlier Factor (LOF). The outlier factor is local in the sense that only a restricted neighborhood of each object is considered [14]. Various other algorithms are proposed for anomaly detection in the Wireless Sensor Networks (WSN). A hierarchical framework have been proposed to overcome challenges in WSN's where an accurate model and the approximated model is made learned at the remote server and sink nodes. An approximated local outlier factor algorithm is also proposed which can be learned at the sink nodes for the detection model in WSN. These provide more efficient and accurate results.

5. Anomaly Detection Methods in Cloud Networks

In cloud networks, traffic or flow of packets comes from more than one domain. There's a rapid change that occurs in the cloud environment due to the patterns or behavior of

clients/tenants using the cloud infrastructure and the state of the unprotected services. In cloud environment, various challenges of identifying anomaly detection such as misconfiguration or high volumes of legitimate traffic in the network. The importance of the anomaly detection in cloud networks is the unwanted activities in data that brings the importance of reason for such anomaly in the information. Generally, the commercial off-the-shelf systems (COTS) for detecting intrusions are based on signatures or rules. Signature based IDS can be used to detect known attacks in the cloud network, although the point of deploy can be before the cloud to detect external or incoming attacks or at the back end of the cloud to detect both external and internal attacks. In the cloud networks, there are different techniques or methods that have been used in the detection of anomalous activities these include:

1. **Statistical Anomaly Detection Systems**
   This method of anomaly detection in cloud base network detects anomaly by observing computations in the network and creates a profile which keeps or stores the generated value in resenting their behavior. In identification of anomaly using this technique, there are two profiles created; the first one stores the normal or anomaly rules or signatures while the second one updates at regular intervals. During the update anomaly scores are calculated. If the threshold value is lower than the current anomaly profile generated, then it is known to be anomalous and detected. There's high probability of occurrence of normal data instances in dense regions of the model, while irregularities is seen in the low possibility regions . Some proposed model of Statistical Anomaly Detection Systems are: Cloud Diag, EbAT (Entropy based Anomaly Testing) etc.The benefits of using this technique are that there is no previous or prior knowledge or training of security risks or knowledge domain required. Additionally, it has the capability of detecting even recent anomaly generated in the network or data and there's accurate notification of anomalies that have occur over extended time frame.
2. **Machine Learning Anomaly Detection Systems**
   The ability for programs or software to improve performance of their task over time by learning is an important technique in the detection of anomaly. Verified values or normal behavior of data are stored, when anomaly occurs or is being detected the machine learns its behavior, stores the new sequence or rule. This technique creates a system that can improve on performance of the program by leaning from the previous results. The interesting part in this technique is that upon improving of performance from previous results, new information are extracted and if it requires a change in the strategy of execution to improve performance it is done on the basis of the new information from the previous results. There are various categories of Machine leaning based anomaly detection such as; Bayesian Network, Genetic Algorithm, Neural Network etc.

   Neural Networks has the capability to improve on data that is not complete to create a potential to detect and understand patterns that are not visible. The Neural network does not only detect previous attacks but also unseen behavior or patterns. Genetic Algorithms employs the evolutionary algorithm techniques such as mutation, selection etc. their different process is based on collected rules from the information on the network analysis carried out by the IDS.
3. **Data Mining Based Anomaly Detection Systems**
   The analyzing or extracting knowledge of large data set to fine patterns that are useful to the data owner is known as Data Mining. This technique uses the classification, clustering and association rule mining methods in the detection of anomalies in cloud environment. An analyst mechanism is in the data mining technique that detects anomaly by differentiating between normal and abnormal activities within the cloud. This is accomplished by stating or delineating some boundaries for valid and normal activities in the cloud network. There is also an added level of focus in this technique for anomaly detection. Data mining techniques are more flexible and easily to deploy at any point. Putting data mining into effect in the cloud network makes available the opportunity to extract meaningful information from data warehouse that are integrated into the cloud, this reduces the infrastructure storage costs. Customers or users of a cloud service only have to pay for the data mining tool that's been used. Data mining is typically used by Cloud Service Providers to provide a much better service for their users or clients using their cloud service. The downside in this is that if the clients are not informed of the information that's been collected and used for mining, there's a violation of their privacy and it's illegal. There are varieties of issues available in data mining detection in cloud based networks which are the priority replacement of preserving privacy and setting the wrong parameters of these privacy settings while using different rules and strategy to enhance cloud network security.
4. **Adaptive Anomaly Detection Systems**
   The Adaptive Anomaly Detection Systems (AAD) employs data description using hyper-sphere for adaptive failure detection. In cloud networks, possible failures or anomaly which are detected by cloud operators are detected by the AAD using the performance data of the cloud service. The AAD detection systems utilize or capitalize on the log of the detected failure records that have been sent in by the cloud operators to identify new types of failures subsequently. The AAD detection algorithm changes its behavior by repeatedly learning from the new certified results or detection from the cloud provider so as to be prepared for future detections. a prototype of AAD system was built and experiment was conducted in it testing the prototype in a 362-node cloud computing environment. It was noted that the prototype was lightweight, and it took couple of seconds to startup the detector and couple of seconds more for the set adaptation and the failure detection to be up and running. 518 metrics were profiled every minute, the profiling covered or circled through the entire statistics of a typical cloud server, its Central Processing Unit usage, task switching processes, memory and swap space utilization, paging and page faults, input and output data transfer, interrupts, and more. Failure detector such as

subspace regularization was used in comparing the ADD algorithm. The failure detector in achieves 67.8% sensitivity in the experiments. The Bayesian sub-models and decision tree classifiers that were proposed only have 72.5% detection sensitivity. In the AAD the failure detector could get up to 92.1% and 83.8% detection sensitivity and detection specificity.

6. Comparison of Survey of Cloud Security Measures in Cloud Data Storage Applications

Cloud storage is a useful way of storing data and also sharing of information online. The important question asked is "is it safe to store sensitive information on the cloud?" well that's a question we are trying to evaluate and answer if possible. Security in the cloud is not all that 100% guarantee. Files maybe encrypted in transmission, and at the final destination, the CSP might decrypt the file to gain access because the encryption algorithm used is provided by them. Access to your account can be gotten by anyone and your sensitive files can be compromised. In this case encryption on the client or the cloud user side is important and also using of a strong encryption key is advised.

Dropbox:
Dropbox is a public cloud storage, which was developed by 2 graduate of MIT who always forget or misplace their USB devices holding information that they need to use momentarily. Due to this Dropbox was brought to light in the IT world. In 2007 Dropbox Inc. was founded, itprovides cloud storage, client software and file synchronization . Dropbox allows it users to upload their files or folders into the Dropbox folder where it can be viewed or shared on any device at any time as long as the device has Dropbox installed along with a username and password and also internet connection for synchronization. Dropbox was developed for personal use that was the intention of the two MIT graduate, but as of 2011 the cloud application have housed over 50 million users worldwide storing over 20 billion files and occupying petabyte of storage. Dropbox gives a 2 GB cloud storage space for free, but additional space can be purchased. Dropbox application is available for windows, Apple OS X, Android, and Linux

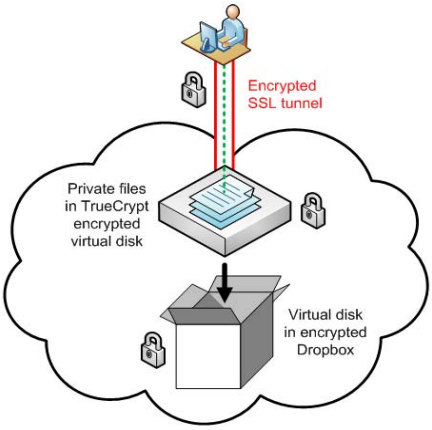

Figure 2: Working Process of Dropbox Protocol

Figure 2 illustrates the example of working mechanism of Dropbox protocol. The basic mechanism is working based on so called hand-shaking process of basic networking standards.

Google Drive:
The "Google Drive" is the Google version of cloud storage, and it is one of the popular cloud services. It supports photos, videos, documents and other files. There's a 15 GB free storage given that can be increased at any time by the user. Google drive provides generic applications for viewing of more than 30 file types without having to install the corresponding application into your computer system for viewing the corresponding file type. The Google drive provides unlimited file size upload quotation for uploading files into corresponding user drive.

iCloud:
iCloud is cloud storage from Apple Inc. It was launched on October 12, 2011 . iCloud offers its users with the means to store data such as; documents, images, videos, etc. users can also backup their iOS devices directly to the iCloud wirelessly. As of July 2013, the iCloud service had 320 million users. The iCloud was first branded as iTools in 2000, Mac in 2002, and MobileMe in 2008.

7. Conclusion

Anomaly detection in could networks is a wide area of research, and it holds a good number of developments and proposing of detection systems. Anomalous activities occur always in our networks cloud based or noncloud based. With the different types of methods or techniques in anomaly detection in cloud based network, detection of unwanted behavior can be traced, detected, stopped. These techniques have their limitations that create a gap between their performance metrics. In cloud based network hybrid anomaly detection system or method should be used so as to have a more efficient and high performance system. In this paper, we have discussed the importance of anomaly detection system in cloud environment, its types, methods, and the limitations that each method is faced with such as, false alarm being created; detection accuracy is hinged on the basis of previous collected information on anomalous behavior; more time is needed in the identification of attacks etc. These limitations can create inaccuracy in anomaly detection. This review will be helpful to researchers for gaining a basic insight of various approaches for the anomaly detection. Although much work had been done using independent algorithms, hybrid approaches are being vastly used as they provide better results and overcome the drawback of one approach over the other. Every day new unknown attacks are witnessed and thus there is a need of those approaches that can detect the unknown behavior in the data set stored, transferred or modified. In this research work fusion or combination of already existing algorithms are mentioned that have been proposed.

References

[1] Chandola V., Banerjee A. , Kumar V., Anomaly detection: A survey, ACM Computing Surveys (CSUR); 41(3); 2009;p. 15 .
[2] Agarwal B., Mittal N., Hybrid Approach for Detection of Anomaly Network Traffic using Data Mining Techniques, Procedia Technology; 6; 2012; p. 996-1003.
[3] Padhy N., Mishra P. , Panigrahi R., The Survey of Data Mining Applications and Feature Scope; International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 2(3) ;2012; p. 43-58.
[4] Lee W., Stolfo J. Salvatore, Data mining approaches for intrusion detection; Proceedings of the 7th USENIX Security Symposium, San Antonio, Texas;1998;p. 79-94.
[5] Lee W., Stolfo S.J., Mok K.W., Adaptive intrusion detection: A data mining approach; Artificial Intelligence Review;14(6);2000; p. 533-567.
[6] Phua C., Lee V., Smith K., Gayler R., A comprehensive survey of data mining-based fraud detection ; research; 2010; p. 1-14.
[7] Chauhan A., Mishra G. , Kumar G. , Survey on Data mining Techniques in Intrusion Detection; International Journal of Scientific & Engineering Research ; 2011; p.1-4.
[8] Xu L., Yeh Y. R., Lee Y. J., Li J., A Hierarchical Framework Using Approximated Local Outlier Factor for Efficient Anomaly Detection; Procedia Computer Science ; 19; 2013; p. 1174-1181.
[9] T. Pang-Ning, M. Steinbach, V. Kumar, Introduction to data mining, Library of Congress, 2006.
[10] Munz,G., Li S., Carle G., Traffic Anomaly Detection Using K-Means Clustering; GI/ITG Workshop MMBnet; 2007;p.1-8. Oliveira, A.C., Chagas, H., Spohn, M., Gomes, R. and Duarte, B.J. (2014) Efficient Network Service Level Agreement Monitoring for Cloud Computing Systems. 2014 IEEE Symposium on Computers and Communications (ISCC), Funchal,23-26 June 2014, 1-6.
[11] Roschke, S., Cheng, F. and Meinel, C. (2009) Intrusion Detection in Cloud. Eight IEEE International Conference on Dependable Automatic and Secure Computing, Liverpool, 729-734.
[12] Zhang, Q., Cheng, L. and Boutaba, R. (2010) Cloud Computing: State-of-the-Art and Research Challenges. Journal of Internet Services and Applications,**1**,7-18.
[13] Wang, C. (2009) Ebat: Online Methods for Detecting Utility Cloud Anomalies. Proceedings of the 6th Middleware Doctoral Symposium, ser. MDS '09. New York, ACM, 4:1-4:6. http://doi.acm.org/10.1145/1659753.1659757.
[14] Hussain, M. (2011) Distributed Cloud Intrusion Detection Model. International Journal of Advanced Science and Technology, **34**, 71-82.

Author Profile

Dr. Chinthagunta Mukundha, Associate Professor, Department of Information Technology, Sreenidhi Institute of Science and Technology, HYD-501301, AP, India.