Journal of Computational Science & Engineering 18 (2015) 605-609

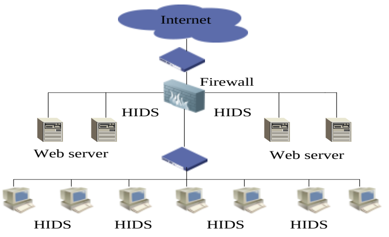
Available at http://www.asocse.org ISSN 1710-4068

Study on intrusion data detection algorithm for user data visa cloud computing

Zhao Tao[1*]

Department of Computer Science, Yunyang Teachers' College, Shiyan, China

Abstract: Cloud computing is a new computing model, it will be large-scale computing resource interconnection were effectively integrated, and the computing resources available to users in the form of services. The user can at any time according to need to access a virtual computer and storage system using a broadband network, without the need to test the underlying anxiety complex implementation and management, greatly reduce the difficulty of realization and hardware investment users. Cloud computing effectively the actual resources and virtual service separation reduce all kinds of business service costs, improve the utilization of network resources. The main work of this paper includes: first introduced the current cloud environment security threat, attack and common intrusion detection technology, summarizes the advantages and disadvantages of the proposed at the present stage of intrusion detection system under the cloud environment. Then the research on the analysis of the fuzzy C mean clustering algorithm for intrusion detection exist deficiencies in cloud environment, the improvement and optimization of its, and the improved algorithm for parallel implementation of map reduce, to solve the clustering problem of the magnanimity data.



Keywords: Cloud computing; Security storage; Symmetric encryption; Intrusion detection

Received 20 April 2015, Revised 22 May 2015, Accepted 25 May 2015

* Corresponding Author: Zhao Tao, doitor2078@sina.com

1. Introduction

Application of security technology existing in cloud facing large-scale data, high concurrency access, service available all-weather, software compatibility and challenges, how to enhance the security of the whole system become the future development trend of intrusion detection and security in cloud service efficiency and quality. Massive data processing and mass calculation is the common problem of intrusion detection, intrusion detection algorithm for many traditional often only suitable for small scale data processing, when the amount of data increases, they tend to increase as the slower or even unable to run the calculation [1]. But cloud computing can greatly compensate for these shortcomings, it provides ultra large scale computing capacity and large storage capacity, can be in the behavioral event collection, correlation analysis, virus prevention and other aspects to achieve significant improvements, but also can be used to construct the test platform of intrusion analysis inspection ultra large scale, enhance the safety situation of whole cloud environments to grasp ability [2].

The emergence of cloud computing can reduce the cost of computer users; make the user experience more performance and unlimited storage capacity. The use of cloud computing, users do not need to worry about creating a user's machine document if the application or operating system with other users of the compatible [3]. When everyone shared data or applications in the cloud, the format is not compatible with the problem no longer exists [4]. Because cloud computing have high availability, easy extensibility and service cost is small, so the access to the vast number of IT users, but the concept of cloud computing in recent years has received wide attention, relevant technology is still not mature enough, has not been widely application [5].

In this paper, in order to satisfy the detection of intrusion cloud environment, deeply studied the current cloud environment faced with various security threats, and common attack means and methods of detection. Through the analysis of the advantages and disadvantages of detection algorithm for intrusion detection system established under cloud environment and invasion, presents the anomaly detection method based on fuzzy C mean clustering in a cloud environment. Optimize and improve the corresponding to the existing fuzzy C means algorithm's shortcomings, and based on the improved algorithm of map reduce based on the realization of parallel. To make the original memory consumption, the amount of calculation algorithm of intrusion detection is very complexes; the invasion by parallel computing framework is applied to large-scale cloud detection, so as to construct an intrusion detection system rapidly and efficiently in a cloud environment.

2. Related theory of intrusion detection

2.1. Research status of intrusion detection algorithm

2.1.1. Research status of traditional intrusion detection algorithm

Traditional intrusion detection algorithm mainly includes three types: misuse detection, anomaly detection and hybrid detection.

Gutirrez [6] provides some misuse detection based on the topic, and finally put forward on the basis of several different intrusion detection framework. The misuse detection in the intrusion detection system can achieve very high accuracy and less error rate based on misuse detection, however, but cannot detect unknown attack. In order to detect new or unknown attacks, intrusion detection algorithm was proposed based on anomaly detection.

Anomaly detection is used to identify with the normal behavior of different events, cannot rely on labeled data sample and detect the unknown attacks. Including data mining, artificial neural network, association rules, and fuzzy algorithm technology has begun to deal with the related problems of anomaly detection.

Dutkevyach proposed a solution based on the real-time intrusion detection anomaly, used to attack and multidimensional data analysis based on network protocol flow. Zhengbing proposed real-time lightweight intrusion detection system; the system uses the behavior pattern and the data mining technology to automatically detect coordinated attacks in the network.

2.2.2. Study status of intrusion detection algorithm for cloud environment

Based on the intrusion detection algorithm on the traditional intrusion detection algorithm, cloud environment have also been studied and developed based on the.

Mazzariello C proposed a network based intrusion detection system in the cloud environment [7], this paper defines a series of rules, to determine the corresponding behavior whether to belong to the intrusion behavior. The system has a higher detection rate, however all intrusion rules need to be predefined, system is not able to detect unknown attacks, has the very high error detection rate for unknown attacks, while the system is very difficult to detect the encryption and virtual network, can only detect external attack.

Put the Bias classifier based on cloud and to the Snort network intrusion detection system application environment. The results show that the system reduces the detection error detection rate and computation cost, and can detect unknown attack [8].

Pardeep Kumar proposed cloud intrusion detection technology of hidden Markov model based on clustering method, with clustering technology to reduce the quantity of the detection, and then using HMM can track the characteristics of state transfer, the establishment of the system of normal program behavior with normal system call sequences of HMM, finally, calculating the probability of its occurrence in normal model for observation sequence, according to the probability size to determine whether the intrusion behavior. Clustering technology to accelerate the speed of detection, HMM can detect unknown attack good any network.

2.2. ***The existing intrusion detection system in cloud environment problem***

The above research are simply to intrusion detection method is applied to a variety of cloud environment detection system, and do not consider the face detection performance when the massive intrusion detection data. When a very large amount of data, the algorithm can perform real-time detection and the detection accuracy rate is very low.

Sun Ya Fang began to consider data analysis log massive intrusion detection using MapReduce algorithm, finally make the calculation speed increases of 89% log analysis. Anna Koiifakou is also proposed for flow analysis using MapReduce, realize in the massive high-speed network traffic statistics [9]. Yeonhee Lee proposed DDos attack detection using Hadoop, solve DDos attack detection in large scale traffic environment. However, so far, the researches on Intrusion Detection System in cloud environment, also cannot really achieve self analysis and real-time detection, but also have good scalability.

Although each a cloud computing provider emphasizes to protect user data using encryption, but even if the data is encrypted, only refers to the data is encrypted transmission on the network, the protection of data in the processing and storage are still unresolved, especially when there is a data storage, because when the data is usually have been declassified, how to protect, it is difficult to solve.

Network intrusion detection system needs to occupy a larger network based resources in robustness, implementation technology have a lot of problems, at the same time for the encrypted network, internal attack, network virtualization are very difficult to detect. Distributed intrusion detection systems are mostly by a central server and a large number of distributions of the local IDS component, wherein the local IDS responsible for local events, the central server is responsible for the overall analysis [10]. The central server may be overloaded, and IDS work in the aspects of centralized distribution is also very difficult to manage, with communication and computation cost is high.

3. Research on intrusion detection clustering method

3.1 . Fuzzy C means algorithm

Based on the assumption that the above problems, establish goals for mixed integer mathematical model to minimize the total completion time:

C means algorithm is executed only in the optimal path on the edge of the standard AS algorithm pheromone update:

$$\tau_{ij} = (1-\varphi)\tau_{ij} + \sum_{k-1}^{m} \Delta\tau_{ij}^{k}$$

The objective function is defined for FCM:

$$J_m(U, c_1, ..., c_c) = \sum\nolimits_{i=1}^{c} J_i = \sum\nolimits_{i=1}^{c} \sum\nolimits_{j}^{n} u_{ij}^{m} d_{ij}^{2}$$

By constructing the following new objective function can be obtained, is the objective function to achieve the necessary condition of the minimum value.

$$J(U,c_1,...,c_c,\lambda_1,...,\lambda_n)=J(U,c_1,...,c_c)+\sum_{j=1}^{n}\lambda_j\left(\sum_{j=1}^{n}u_{ij}^m-1\right)$$

For all the input parameters for the partial derivative, which make the objective function achieve minimal prerequisites:

$$v_i=\frac{\sum_{j=1}^{n}u_{ij}^m x_j}{\sum_{j=1}^{n}u_{ij}^m}$$

Fuzzy C means clustering algorithm is a simple iterative procedure, FCM according to the following steps to determine the cluster centers and membership matrix S:

$$\begin{aligned}
&\sum_{j=1}^{a}X_{jb}\le B && b=1,\dots,k\\
&P^b\ge X_{jb}p_j && j=1,\dots,n;b=1,\dots,k\\
&S^b\ge X_{jb}\tau_j && j=1,\dots,n;b=1,\dots,k\\
&S^b\ge S^{b-1}+P^{b-1} && b=2,\dots,k\\
&X_{jb}\in\{0,1\} && j=1,\dots,n;b=1,\dots,k\\
&C_{\max}\ge S^b+P^b && b=1,\dots,k\\
&\left\lceil\frac{n}{B}\right\rceil\le k\le n
\end{aligned}$$

3.2. Throwing type fuzzy C means algorithm

When the data with uncertainty will be difficult to cluster analysis was performed by using common method, must be in the proposed fault-tolerant mechanism based on clustering. We introduce a new distance metric distance one fault tolerance, to keep in the cluster in a reasonable amount of similarity between objects, and can deal with noise problems of the huge data. The N dimension between the data objects and cluster is poor as its fault tolerance distance, as shown in the formula:

$$d_t=\left\|x_i-v_j-\left|x_i-v_j\right|\right\|$$

In the objective function is defined fault distance the new metric for FCM algorithm:

$$J=\sum_{i=1}^{n}\sum_{j=1}^{c}u_{ij}^m\left\|x_i-v_j-\left|x_i-v_j\right|\right\|^2$$

The objective function we obtain:

$$J_{ME}=+\sum_{i=1}^{n}\frac{x_i}{N}\sum_{i=1}^{n}\sum_{j=1}^{c}u_{ij}$$

3.3. Intrusion detection system based on cloud model

Cloud system is divided into three layers: system layer, platform layer, application layer. Figure 1 shows the system model in cloud environment, where the system layer includes a virtual host and network, such as the Amazon EC2 service, it provides a virtual host and network service. The platform layer is the second layer of cloud system, including the operating system virtualization and running environment and APIs, such as Windows Azure, it provides some APIs, used to store and manage some independent of the common language runtime environment (CLR). The application layer as the top of the cloud system, responsible for providing virtual applications, such as Google App Engine, the customer can write and upload some Web application; they can be executed on the GAE and can be accessed via web.

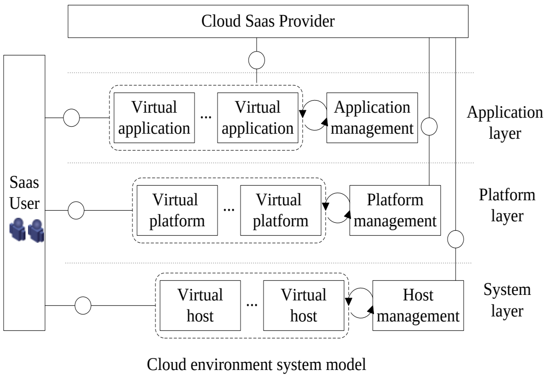


Figure 1. The system model in cloud environment.

Host intrusion detection system from the host monitoring and analysis of the collected information based on the specified. HIDS use the collected information, such as file systems, networking events, and system call to detect intrusion information. Through the observation of the host kernel modifications to the host file system and the behavior of the program, when detected with expected behavior does not meet the report of possible attacks. The efficiency of HIDS depends on the characteristics of the monitoring system was shown in Figure 2.

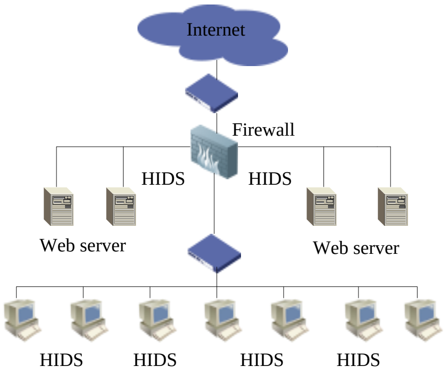


Figure 2. The efficiency of HIDS depends on the characteristics of the monitoring system.

3.4. The architecture of HDFS

HDFS uses M/S architecture, each HDFS cluster is composed of a name node (Namenode) data nodes and a certain number of (Datanode), as shown in Figure 3. HDFS in the underlying implementation, first of all is the document is cut into pieces, then these blocks are stored in the data nodes distributed on different. Each block is copied into a plurality of parts, stored in different data nodes, in order to achieve the purpose of fault tolerance. The management node is a central server, is responsible for the management of the file system name space (Namespace) and the file access client. Data node in the cluster is responsible for the management of it lies on the node storage. NameNode is the kernel of HDFS, it is through the maintenance of a group of data structure to record each file is cut into many pieces, these blocks can be obtained, from which the data node in each data node state and other important information.

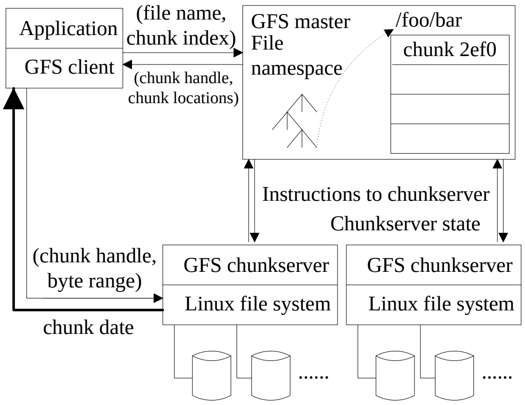


Figure 3. HDFS cluster is composed of a name node (Namenode) data nodes.

3.5 The MapReude programming model

Processing flow of MapReduce operation is divided into 3 steps to perform, respectively (Map phase decomposition parallel tasks (Combine stage), improve the efficiency of the reduce (Reduce), the stage of processing results summary). MapReduce distributed programming model of each stage of processing flow was shown in Figure 4 based on Hadoop:

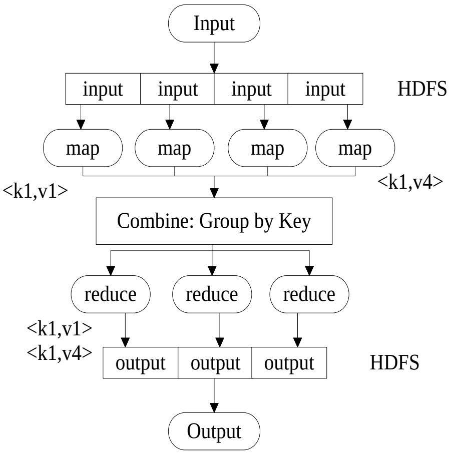


Figure 4. MapReduce distributed programming model of each stage of processing flow.

4. The experimental analysis

In order to verify the performance of PLSVD algorithm to reduce the dimensions for massive high dimensional intrusion detection data, 500000 records we choose KDDCUP 99 data set to test, for the 41 characteristics of CUP 99 data set KDD amount, when the data dimension reduction by using the traditional PCA algorithm, the memory overflow phenomenon will occur. Here we use a parallel PCA algorithm and PLSVD algorithm to reduce the dimension of data contrast verification. Obtained in the experiment, when the expected degradation dimension is 10, first 10 singular value and occupy the singular value above 85%, can meet the requirements of the experiment. So select the dimension from the beginning of 10, the two algorithm time overhead as shown in Figure 6. Experiments show that, in the condition records under section 500000, when choosing the different dimensions of data dimensionality reduction, PLSVD time cost of smaller, more suitable for the requirement of real-time intrusion detection.

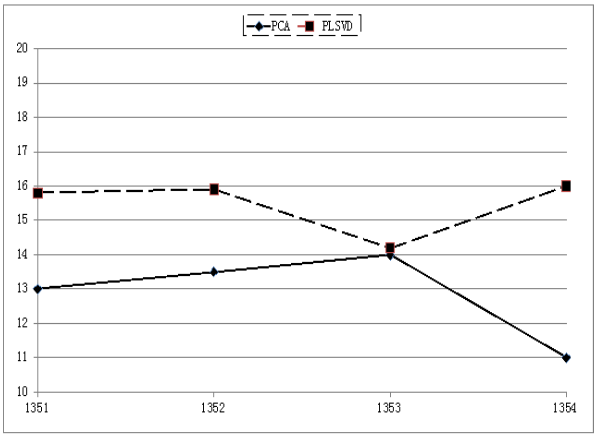


Figure 5. The two algorithm time overhead.

5. Conclusion

According to the result of analysis the threat model, the sensitive data in the network transmission, and storage to the cloud database may have security problems, the sensitive data specific to the user, encryption processing optimization, using a different account according to the specific user using a symmetric key are different in information transmission and storage to the database of symmetric encryption, asymmetric encryption double encryption methods used and the symmetric key in the process of transmission.

In this paper, through the research and improvement of detection algorithm related to invasion, and applies the improved algorithm parallelization deployed to the cloud environment, committed to the establishment of a in a cloud environment based on distributed unsupervised learning of intrusion detection system. The system can real-time detection of autonomous learning, intrusion detection of massive data, and be able to automatic analysis of unknown attacks and recognition, has high detection rate, low false detection rate, and has the advantages of high performance concurrent query engine and good scalability.

References

[1] T. Dutkevych, A. Piskozub, N. Tymoshyk, IEEE, (2013) 599-602.

[2] M.C. Kim, J. Park, W. Jung, H. Kim, Yo.J.

Kim, Annals of Nuclear Energy, 37 (2010) 888-893.

[3] A. Radonjie, V. Vujicie, Information Processing Letters, 110 (2010) 518-520.

[4] S.W. Kim, J. Park, S.Y. Han, H. Kim. Journal of Loss Prevention in the ProcessIndustries, 23 (2010) 539-548.

[5] A. Vacearo, D. Villacei, Electric Power Systems Researeh, 73 (2005) 287-294.

[6] B.F. Gutirrez, S.S. Ben, J. Pijoan, Adaptive resource management for a MC-CDMA system with mixed QOS classes using a cross layer strategy.

[7] C.H. Lien, H.C. Chen, Y.W. Bai, M.B. Lin, Instrumentation and Measurement Technology Conference proccedings, (2008) 2-79.

[8] L. Xu, T.D. Chen, Z.G. Ren, D. Wu, Global Mobile Congress, (2007) 284-288.

[9] M. Khodier, G. Saleh, International Journal of Electronics and Communications, 64 (2010) 489-502.

[10] Y.J. Chung, C.H. Paik, H.G. Kim, First International Conference on Communications and Elcetronies, (2007) 5.