Applied Mathematics & Information Sciences
An International Journal

Data Security Strategy Based on Artificial Immune Algorithm for Cloud Computing

Chen Jinyin, Yang Dongyong

College of Information Engineering, Zhejiang University of Technology, 310023 Hangzhou, China

Received: 15 Oct. 2012, Revised: 27 Nov. 2012, Accepted: 6 Dec. 2012
Published online: 1 Feb. 2013

Abstract: With the fast development of cloud computing and its wide application, data security plays an important role in cloud computing. This paper brought up a novel data security strategy based on artificial immune algorithm on architecture of HDFS for cloud computing. Firstly, we explained the main factors influence data security in cloud environment. Then we introduce HDFS architecture, data security model and put forward an improved security model for cloud computing. In the third section, artificial immune algorithm related with negative selection and dynamic selection algorithm that adopted in our system and how they applied to cloud computing are depicted in detail. Finally simulations are taken by two steps. Former simulations are carried out to prove the performance of artificial immune algorithm brought up in this paper, the latter simulation are running on Cloudsim platform to testify that data security strategy based on artificial immune algorithm for cloud computing is efficient.

Keywords: Data security, Artificial immune, Cloud computing, HDFS, Cloudsim

1. Introduction

Cloud computing is a new computing paradigm appeared in 2006, and the evolutionary offspring of parallel computing, distributed computing, utility computing and grid computing, and the developmental outcome of network storage, virtualization and load balance[1–3]. The main idea of cloud computing is to build a virtualized computing resource pool by centralizing abundant computing resources connected with network and present the service of infrastructure, platform and software. This network that offers various computing resources is called cloud [4–7]. As a supercomputing paradigm based on the Internet, cloud computing allows customers to dynamically share a mass of hardware, software and data resource, and charges according to their actual usage. Therefore, computing power can be sold and purchased as merchandise easily by network in a low price, just like water, gas and electric power [8,9]. Cloud computing is an innovatory thing similar to electric power changing from a single generator to a centralized electric power plant.

Cloud computing has been encountered with security problems. In this paper we want to carry out security strategy for cloud computing, the rest of the paper is organized as follows: We present the data security problem of cloud computing in the next section and then discuss the details of requirement of security in Section 2. In Section 3, we focus on the Data security strategy based on artificial immune algorithm. Then, we get experimental evaluation for the proposed strategy for cloud computing on CloudSim platform and analyze the performance in detail in section 4 and section 5. Finally, we conclude the paper in last section.

2. Data security problem of cloud computing

2.1. Data Security Problem

Compared with traditional software architecture, cloud computing has more serious data security problem. Data security is aimed at applying technical mechanism to guarantee data management in reasonable control, and guarantee data without illegal visit or revise during data process. As mentioned above, the main data security problems of cloud computing could be summarized as follows.

(1) Data security of virtual machine. Virtualization is the kernel of cloud computing, while virtual machine is the most important part of virtualization, which makes

[*] Corresponding author e-mail: chenjinyin@163.com

data security especially important for cloud computing. Current virtual machine of cloud computing has security threats, so how to protect virtual machine with back holes makes great meaning for cloud computing. Because the application network is quite complex, novel methods are applied to solve data security problem of virtual machine, including artificial intelligence, pattern recognition, artificial immune et al. In this paper, we brought up a novel artificial immune mechanism based data security strategy for virtual machine.

(2) Data migration security. For cloud computing, data are stored in cloud, which will also be backed up in several places which make data security significant. This is the act of moving data from one location to another. In dynamic cloud environment, data migration should be taken into consideration in prior.

(3) Data management security. Data management is the key work for cloud computing provider, because they have to deal with consequences of improper data management or unreasonable data store structure. A complete data management mechanism plays a major role in data security for cloud computing.

2.2. HDFS architecture

Hadoop is the next open source basic software project of Apache, which gains wide applications nowadays. As basic component of Hadoop, HDFS (hadoop distributed file system) has data redundancy capability in application which could create developments for data security strategy. HDFS is used in large-scale cloud computing in typical distributed file system architecture, its design goal is to run on commercial hardware, due to the support of Google, and the advantages of open source, it has been applied in the basis of cloud facilities. HDFS is very similar to the existing distributed file system, such as GFS (Google File System). They have the same objectives, performances, availability and stability. The main architecture of HDFS is shown in following figure.

HDFS has a master/slave architecture. An HDFS cluster consists of a single NameNode, a master server that manages the file system namespace and regulates access to files by clients. In addition, there are a number of DataNodes, usually one per node in the cluster, which manage storage attached to the nodes that they run on. HDFS exposes a file system namespace and allows user data to be stored in files. Internally, a file is split into one or more blocks and these blocks are stored in a set of DataNodes. The NameNode executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes. The DataNodes are responsible for serving read and write requests from the file systems clients. The DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode.

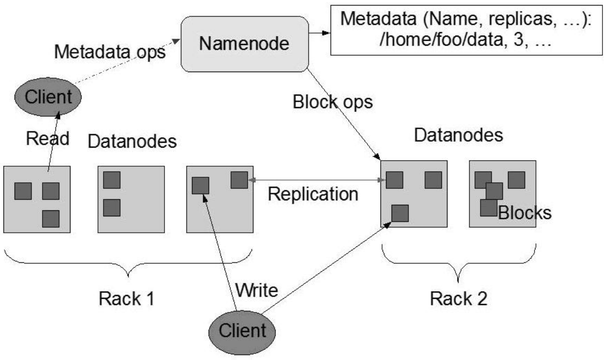

Figure 1 HDFS architecture

For HDFS architecture, three potential threats exist in data applications, including four-level structures: authentication, public-key cryptography, privacy protection and file recovery. Authentication is responsible for users log in control, which could protect privacy by digital certification and authorities managements. Public-key cryptography is one traditional way to guarantee key management from hiker intrusion. Privacy protection exists if NameNode is attacked or failure, there will be disastrous consequences on the system. So the effectiveness of NameNode in cloud computing and its efficiency is key to the success of data protection, so to enhance NameNodes security is very important. The rapid recovery of data blocks and r/w rights control: Datanode is a data storage node, there is the possibility of failure and can not guarantee them availability of data. Currently each data storage block in HDFS has at least 3 replicas, which is HDFSs backup strategy. When comes to how to ensure the safety of reading and writing data, HDFS has not made any detailed explanation, so the needs to ensure rapid recovery and to make reading and writing data operation fully controllable can not be ignored.

3. Data Security Model

Classical cloud computing technology could be described as following mathematical model.

$$D_k = C(NameNode) \quad (1)$$

$$K_k = f \times D_f \quad (2)$$

where $C(.)$is short for node's visit, $NameNode$ in Eq. (1) represents system application servers. $D_f$donates chunk file matrix, $K_f$indicates chunk file matrix for data center of system. $f$is file itself, file f in system could be shown as $f = F(1),F(2),...,F(n)$which means file f has n chunks, $F(i) \cap F(j) = \oslash, i \neq j, i, j \in 1,2,3,...,n$. To enhance the data security of cloud computing, we provide a Cloud Computing Data Security Model called C2DSM.It can be described as follows:

$$D'_f = C_A(NameNode) \tag{3}$$

$$D_f = M \times D'_f \tag{4}$$

$$K_k = E(f) \times D_f \tag{5}$$

where $C_A(.)$: authentic visit to $NameNode$; D'_f: private protect model of file distributed matrix; M: resolve private matrix; $E(f)$: encrypted file f block by block, get the encrypted file vector.

4. Data security strategy based on artificial immune algorithm

4.1. Problem description

In cloud computing environment, several reasons could threat data security including uncertain cloud network, network node broken down, super-large user visits and data unconsistency. One of the above situations would make data insecure. We maintain data consistency to keep data block on each node addressable. So maintain system consistency is to guarantee viewgraph consistency instead of data addressable on each online node.

We suppose a user wants to visit current data task could be presented as $f = f_1, f_2, ., f_n$. While visiting file f, n task will be assigned on k data nodes. Each node runs task T_i, so the feasibility of f can be donated as follows.

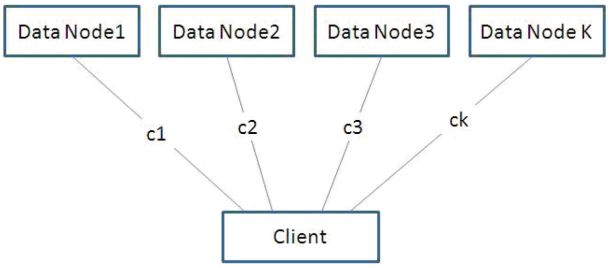

Figure 2 Data visit feasibility model [8]

where C_i indicates the network connection between client and data node i. If and only if data node f_i and C_i are both addressable, then file f could be obtained. So the feasibility of file f could be described as

$$f = \prod_{i=1}^{k} C_i f_i \tag{6}$$

4.2. Data security strategy based on artificial immune

In order to keep data consistency during file visit, a novel data security strategy is brought up to manage data and file store. Two aspects need to consider for file store and management in cloud computing.

1. Data block numbers. Number of data block is used for count data block numbers stored in the same physical node site. The data creator needs to consider how many blocks should be created, the more blocks are, the more resources will be wasted which will cause more data consistency maintain cost. The less block number is, the data block cannot meet the demand of user visit. So the data block number effect the data management and consistency maintain.

2. Data block granularity. Data block granularity decides the file system efficiency by storing different files. Data block of HDFS is 64M, while other file may adopt different block granularity. File creation and management could be simulated as artificial immune system. Following table shows their common ground.

File stored in cloud environment, they firstly be coded as antibody by using binary coding rules.

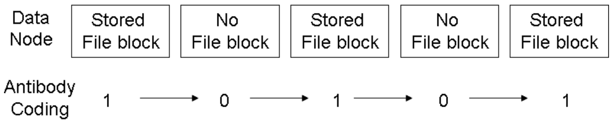

Figure 3 Antibody coding rule of stored files for cloud computing

For instance file stored in node as above figure, HDFS has five data nodes, "1" is donated for file has already stored in the data node, while "0" represents the file has lost or cannot be addressed. In Hamming rules, antigens and antibodies are represented as sequences of symbols. The mapping between sequence and shape is not fully understood, but in the context of artificial immune systems, they are assumed to be equivalent. Following depicts the Hamming distance measure.

$$D = \sum_{i=1}^{L} \delta, where \delta = \begin{cases} 0 \text{ when } ab_i \neq ag_i \\ 1 \text{ when } otherwise. \end{cases} \tag{7}$$

Hamming distance is only applied for binary-coded antibody and antigen, which is suitable for file coding situation

The goal of this algorithm is to cover the non-self space with hyper-spherical antibodies. Specifically, it can produce a good estimate of the optimal number antibodies need to cover the antigen, and maximization of the non-self coverage is done through an optimization algorithm with proved convergence properties. The algorithm is based on a type of randomized algorithms called Monte Carlo methods. Specifically, it uses Monte Carlo integration and simulated annealing. In our work, negative selection algorithm is adopted to mature antibody while files stored in data node could be detected and arranged in optimized node. We suppose the file bock number is k. Number of antibody is k has been generated randomly and matured through negative selection algorithm. The antibody set is $D = d_1, d_2, d_n$.According to matching function defined by Eq. (7) , each antibody's fitness can be achieved and ordered. Clone selection is main strategy for new file create and management into the data node. So dynamic selection algorithm is adopted to manage add new file into the data node depicted in next section.

Kim and Bentley adopt such a strategy as a clone selection operator with negative selection operator for network intrusion detection. They conclude that the embedded negative selection operator plays an important role. The main idea is that for those valid detectors generated, if a bit or several bits are changed, their fitness score will not vary in a large extent. Thus more valid detectors will be obtained in a short time. The steps are shown as follows. Step 1. System initialization. Initialize data node of system with $(D_1, D_2, ..D_n)$, and current file visit is denoted as $a_i, i = 1, 2.n$; Step 2. Initialize data node set S1, if $D_i \subset S_i$, then the data visit capacity is $a_i = a_i + 1$. If user ends the D_i visit, then visit capacity of D_i will be donated as$a_i = a_i - 1$. Step 3. Locate the file, and visit the data node set S_k. If $D_i \subset S_k$, then visit capacity of node D_i is $a_i = a_i + 1$. If user ends the visit of D_i, then the visit capacity of node D_i is $a_i = a_i - 1$. Step 4. If the data node locates on the same data block, then select the minimized a_j as the visit node for user to access is the optimized management.

5. Experimental evaluation

5.1. Immune algorithm simulation on benchmark functions

Before we apply artificial immune algorithm to data management of cloud computing, we firstly test the performance by optimizing benchmark functions: Sphere, Rosenbrock, Gastrigin and Griewangks.The simulation figures and comparison tables testify that artificial immune algorithm achieves better optimum on the four benchmark functions. When optimizing Sphere function, immune algorithm arrives at precision of 1E-59 which approaches zero. Especially in case of Griewanks optimum value of immune algorithm is much better than other evolutionary algorithms. The optimized results and comparison are shown as follows.

Artificial immune algorithm is brought up to optimize multiple dimension functions in this section. Simulation results testify that artificial immune algorithm can attain better optimum, which is the basis for its application for data optimized store in cloud computing.

Table 1 Simulation results of GA, PSO and improved J strategy combined DQPSO optimizing on benchmark functions

Problem	*GA*	*PSO*	*SA*	*IA*
Sphere	0.0415	0.021	0.012	0.023
Rosenbrock	15.233	14.31	13.67	3.21
Rastrigin	12.31	10.76	1.21	0
Griewanks	1.2E-4	4.23E-4	4.2E-2	2.1E-6

5.2. Simulation of data security strategy based on artificial immune algorithm

The simulation was run on CloudSim platform, based on which a cloud computing data security simulation system are implemented by C++ in this paper. HDFS system consists of 100 machines and one master sever. Each machine adopts 500 disk space and 4G memory. HDFS file store system counts for the data center of cloud computing system. In the simulation the data management system employs two working models: data security strategy on and off. Without security strategy, the file number and status are shown as follows.

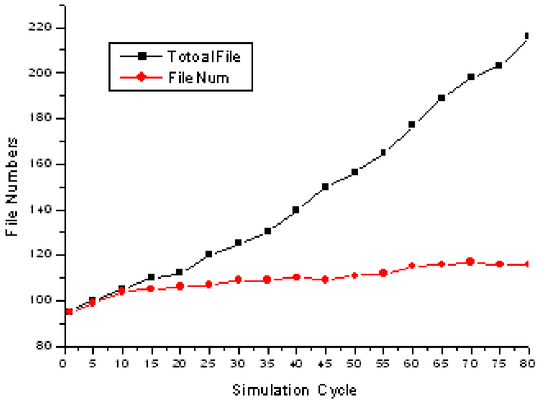

Figure 4 System file status without data security strategy

6. Conclusion

As the development of cloud computing, security issue has become a top priority. This paper discusses the cloud

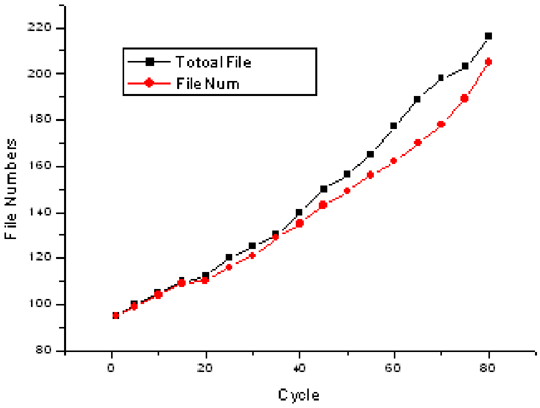

Figure 5 System file status with data security strategy based on AI

computing environment with the safety issues through analyzing a cloud computing framework–HDFS's security needs. Finally we conclude a cloud computing model for data security.

Acknowledgement

The first author acknowledges the financial support by the Natural Science Foundation of Zhejiang Province No.Y106735 and No. Y1100378
The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] Rajkumar Buyya Market-Oriented Cloud Computing:Vision,Hype,and Reality for Delivering IT Services as Computing Utilities (2008), 322-326.
[2] Jean-Daniel Cryans, Criteria to Compare Cloud Computing with Current Database Technology (2008), 23-26.
[3] Shun-Hung Tsai, Y.-H. Chang, Sliding Mode Control for A Class of Uncertain Time-Delay System, Applied Mathematics & Information Sciences. (2012), 53S-59S.
[4] Mladen A. Vouk Cloud Computing Issues, Research and Implementations Journal of Computing and Information Technology - CIT 16, (2008), 235246
[5] Huantong Geng, Yanhong Huang, Jun Gao and Haifeng Zhu. A Self-guided Particle Swarm Optimization with Independent Dynamic Inertia Weights Setting on Each Particle. Applied Mathematics & Information Sciences. (2012), 31S-34S.
[6] Cloud Computing Security: making Virtual Machines Cloud-Ready, www.cloudreadysecurity.com (2008), 34-45.
[7] Dai Yuefa, Wu Bo, Gu Yaqiang, Zhang Quan, Tang Chaojing, "Data Security Model for Cloud Computing", Proceedings of the 2009 International Workshop on Information Security and Application (IWISA 2009), 141-144.
[8] Dai Yuefa, Wu Bo, Gu Yaqiang, Zhang Quan, Tang Chaojing, "Data Security Model for Cloud Computing", Proceedings of the 2009 International Workshop on Information Security and Application (IWISA 2009), 141-144.
[9] David Chappell ,Introducing the Azure Services Platform October (2008), 1232-1240.

Chen Jinyin received the PhD degree in Control Engineering and Control Theory at Zhejiang University of Technology, Hangzhou, China. Her research interests are in the areas of applied intelligent algorithm and network security for cloud computing.