Research and implementation on cloud computing security based on HDFS

Zhilong Liu

School of Mathematics and Computer Science, XinYu University, JiangXi, 338004, China

Corresponding author's e-mail: liuzhilongxyu@163.com

Received 25 October 2013, www.cmnt.lv

Abstract

This paper focuses on the research of the cloud computing security, proposing the file data management model and implementing the security of the cloud computing based on HDFS. The design of the file data management system under the cloud computer environment is achieved based on HDFS, which is with the functions of upload and download data parallelism, user management, inventory management, etc. Aimed at solving the disadvantages of HDFS that is lack of security storage and transmission of the file data, and the integrity checking, the access authentication security mechanism of data node to the client under HDFS, which is based on the IBE algorithm.

Keywords: cloud computing; identity authentication; cloud security; HDFS; access control; safe storage.

1 Introduction

Cloud computing is TCP/IP based high development and integrations of computer technologies such as fast micro processor, huge memory, high-speed network and reliable. Cloud Computing emerged as an effective reuse paradigm, where hardware and software resources are delivered as a service through Internet [1]. Cloud computing is an emerging and increasingly popular computing paradigm, which provides the users massive computing, storage, and software resources in demand [2]. As cloud computing is mature gradually, cloud storage develops rapidly. Using cluster application technology, the grid technology and distributed parallel processing technology, cloud storage works by a lot of different types of storage equipment, providing a dynamic and expanded storage service. Cloud computing is currently getting considerable attention in both academic and industrial areas. With more and more cloud application being available, data security becomes an important issue in cloud computing. Data protection is a critical issue in cloud computing environments. Some studies on distributed storage systems and architectures used in cloud computing environments can be found in [3, 4]. In HDFS [5], all files are stored in text and controlled by a central server. Thus, HDFS is not secure against storage servers that may peep at data content. Additionally, Hadoop and HDFS have a weak security model. In particular the communication between data nodes and between clients and data nodes is not encrypted.

The convenient access for shared configurable computing resources (network, servers, storage, applications and services, etc.) and on-demand access for resources are realized [6]. These resources can be quickly prepared and used as required by tiny management cost or interactions with the service provider. Cloud services can be divided into three layers: infrastructure as a service (IaaS), platform as a service (PaaS) and software as a service (SaaS).The basic principle of cloud computing is that through the network, the huge computing program automatically can be divided into many smaller subroutine, then the large system that are composed of a number of servers, sends the results back to the user after searching, calculating and analyzing [7]. Through this technology, service providers can deal with terabytes of data in a few seconds on the far side, which can achieve the same powerful performance or even better of network with super computer service [8].

2 The relevant theory of HDFS and the cloud computing security

HDFS (Hadoop Distributed File System) is one of the subprojects in the open source project Hadoop, which is under the Apache. Based on the GFS (Google File System) which adapts the access mode of data flow, the functions of storage for large files and big data can be made full use of. The system is equipped with the performance of high throughput and easy deployment.

2.1 THE SYSTEM ARCHITECTURE OF HDFS

The main-slave structure is adapted by HDFS. And the HDFS cluster mainly consists of one name node and several data node. Additionally, one secondary name node may be also included which will periodically communicate with the name node and backup the metadata information in name node. And name node is the primary server of HDFS which takes full charge of the metadata management of the distributed file system and is also responsible for the file access operations of clients. But it is actually not responsible for the file storage. Its' function is equivalent to the file system's control center. Data node is responsible for the storage of data block and response to the request of the client, completing read-write operations for the file data. The system architecture of HDFS is shown in figure 1.

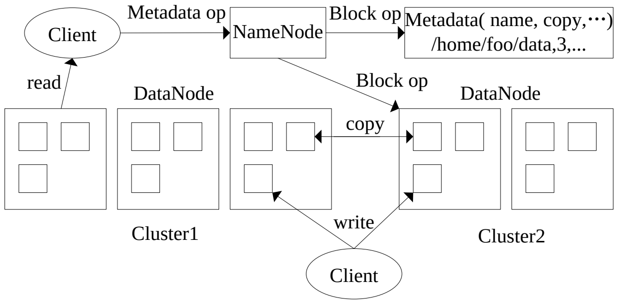

FIGURE 1 The system architecture of HDFS

The name node of NDFS controls all the operations related to read-write. All requests will be processed by name node so that the request side will get right information. According to the type of the server, the client or data node will actually complete the rest service functions. To keep the consistency of data, other operations like read will not be allowed by mutually exclusive operation when written by name node.

2.2 INTRODUCTION OF CLOUD COMPUTING

The resource of cloud computing can be extended with no limitation and gotten based on requirement. The cloud computing mainly involves Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS) and other cloud computing security content.

Cloud computing is different from the traditional mode of the Internet. The satisfactory service can be gotten by only a few equipments. The degree of user’s control over the computing resources is greatly reduced. Among the cloud computing applications, users will upload a lot of sensitive data to the cloud side provided by cloud server providers. The un-trusted factor exists between the user and the cloud service provider. The user can't completely trust the cloud service provider for internally illegal use of private data. The cloud service providers could not keep the user's identification from leaking or faking, resulting in losing users’ privacy and sensitive data. The cloud computing security is important for both users and the cloud service providers. The cloud computing security has following requirements such as the data transmission security, data storage security, the data purge and the application security.

The login of terminal user is mainly solved by the identification of the system. Nowadays, the mainstream identification technology involves PKI which is based on the public key infrastructure and the IBE that is based on the Elliptic Curve Algorithm. The users’ access control is achieved by VPN technology based on SSL. And data access control technology is on the premise of completing the user identity authentication. According to certain rules of permissions, users are divided into different access levels. And different levels have different access. The current mainstream technologies in the access control are mainly DAC (discretionnary access control), MAC (mandatory access control) and RBAC (role-based access control).

As a new computing mode and new technology, Cloud computing, it is the development of grid computing, parallel computing and distributed computing, as well as the new technology of next generation of Internet and application. Resource Scheduling Policy in Cloud computing is an important part of cloud computing technology, it mainly focus on how to allocate compute nodes for the task submitted by users, how to carry on the dynamic extension of the compute nodes in the case of meeting the requirements of service quality from customers and taking the shortest execution time to create the highest degree of load balancing, and its efficiency directly affects the performance of the entire cloud computing environment [9].

3 Design and implement for file management system based on HDFS

Cloud computing systems have similarities and differences with general server systems. To some extent, both of them are processing services and the management center of controlling service. The difference is that the cloud computing platform consists of a large amount of service nodes which have different functions. The cloud computing system generally consists of the main service control cluster, application service cluster, data processing service cluster, data storage cluster and terminal equipments.

In this paper, the design of cloud computing file management system is based on the HDFS whose core idea is derived from the principle and architecture of HDFS. Based on the characteristics of HDFS, the system consists of three parts: the cluster of main service control center, storage service cluster and the client. And storage service cluster is composed of data nodes which are equipped with the characteristics of large quantities and wide distribution. It is the data storage center in the file system with large storage capacity. The data nodes are mainly responsible for processing the data real access operations. And the storage services will not do deep processing for the data. Only according to the command of the name node and the users' requests for data access operations, vast amounts of storage space are provided to ensure the reliability and availability of data [10].

The system architecture is shown in figure 2. The whole system is divided into three parts. The client side is mainly responsible for interaction with users that is the important judgment of the system user experience. All kinds of users’ service requests are directly delivered to the master service by the client side. Through response information of the control center, the file data is fragmented. According to the settings of the block size, the data is copied to the corresponding storage nodes based on the redundancy strategy of HDFS data block. The block data distributed in different data nodes in download system are merged into complete file through metadata mapping information provided by the control servers.

The name node is responsible for processing all kinds of service requests from users, updating the metadata mapping information for MySQL and response from the storage nodes. The name node is the core of the whole system, which has interaction with every module in the system.

In this paper, MySQL database is taken into use to store the metadata information in the file system. Although the user experience of database operation is not as good as that of operating memory directly, the reliability of data can be ensured largely. In this way, disastrous consequences may not be caused by the outages and other accidents. At the same time, the log maintenance work will also be reduced.

According to the practical experience and architecture design, the database method is feasible.

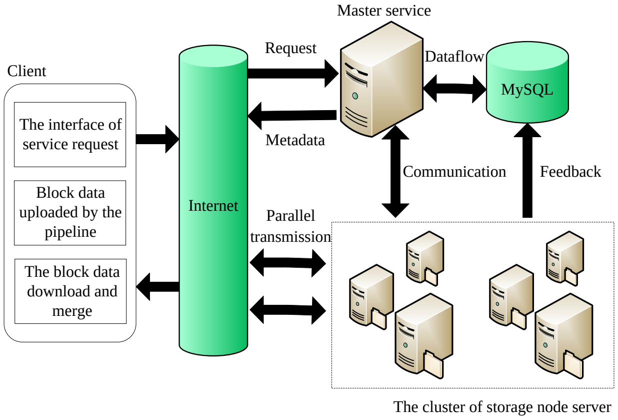
Master service
Client
Request
Dataflow
MySQL
The interface of service request
Metadata
Block data uploaded by the pipeline
Communication
Feedback
Internet
Parallel transmission
The block data download and merge
The cluster of storage node server

FIGURE 2 The system architecture

4 Design of the cloud computing security scheme and optimization algorithm

On the premise of realizing the system function, user audit management, the client identity authentication, data storage, system log processing and access control are involved in the construction. The availability, integrity and reliability of the system can be ensured in the cloud computing file management.

4.1 THE USER AUDIT MANAGEMENT

On the premise of realizing the system function, user audit management, the client identity authentication, data storage, system log processing and access control are involved in the construction. The availability, integrity and reliability of the system can be ensured in the cloud computing file management.

4.2 IDENTITY AUTHENTICATION OF DATA NODE

When the user requests for access to the file data, the metadata information about the files is required. Then the name node returns meta-information. Based on the meta-information, the clients independently communicate with the storage nodes by data links. A fatal security hidden danger exists in this strategy. Malicious illegal users can achieve the metadata information from the complex network environment without the interactions with the name node. But the relevant data is stolen by communicating with data node directly that may bring irreparable damage to the users. In order to prevent the security hidden danger, the client identity authentication will be done by the data node to ensure the security of client's access to name node. The identity authentication scheme is realized in this paper mainly based on IBE system.

4.3 THE OPTIMIZATION ALGORITHM

In the initialization stage of the ant colony algorithm, the pheromone on the nodes of the initial solution is strengthened a certain multiple based on particle swarm optimization (PSO) algorithm for scheduling results, which can makes convergence of ant colony algorithm accelerating. The ants can quickly close to the optimal solution in the process of travelling. Through the many times of experiments, the results show that 6 times pheromone is best.

Some ants are randomly placed on the nodes to travel. The transition probability of node x_{ij} for the first k ant in the time of t are as follows:

$$p_{ij}^{k}(t)=\begin{cases}\dfrac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}(t)]^{\beta}}{\sum\limits_{xis\notin tabu_k}[\tau_{is}(t)]^{\alpha}[\eta_{is}(t)]^{\beta}}, x_{ij}\notin tabu_k \\ 0, \text{else}\end{cases}. \tag{1}$$

$$\eta_{ij}(t)=\frac{1}{CT_{ij}^{k}(t)}. \tag{2}$$

When the ant selects a node, the task was assigned to some virtual resource. And then the pheromone of node will be changed. According to the expression (3) and expression (4), the pheromone can be updated locally.

$$\tau_{ij}(t+1)=(1-p)\cdot\tau_{ij}(t)+\sum_{k=1}^{m}\Delta\tau_{ij}^{k}(t). \tag{3}$$

$$\Delta\tau_{ij}^{k}(t)=\begin{cases}\dfrac{Q}{CT_{ij}^{k}(t)} \\ 0\end{cases}. \tag{4}$$

Among $\rho\in[0,1)$ is pheromone volatilization coefficient. $1-\rho$ represents the information remaining coefficient. $\Delta\tau_{ij}^{k}(t)$ means the information amount remained in the path (i,j) and Q is a constant.

Among the expression, we have

$$x_{ij}\in\{0,1\},\sum_{i=1}^{m}x_{ij}=1, i\in\{1,2,\cdots,m\}, j\in\{1,2,\cdots,n\}$$

5 Experiment and result analysis

After completing the system design and development, through the deployment of test environment, the main functions of the file management system based on HDFS cloud computing is to be tested by test experiments and system CPU power test experiments.

5.1 THE DEPLOYMENT OF THE TEST ENVIRONMENT

HDFS cluster server is set up as follows: ten of the personal computers are taken into use. One of them is regarded as the primary server name node. And the remaining nine personal computers are treated as the data nodes which are responsible for data storage. The Java language is adopted in this test system development. Each PC needs to equip with the JDK whose version is java 6. Keeping the stability and operation efficiency of the system, the system is running on Linux which is for the sake of the safety and reliability of the system's internal communication. The internal secure communication in the cluster is achieved by protocols of Open-SSH.

The software version of the module is shown in table 1.

TABLE 1 The software version

Software	Version
Ubuntu Linux	Ubuntu-12.10-desktop-amd64
Apache tomcat	Apache-tomcat-6.0.16
MySQL	Mysql-essential-5.1.46
JDK	Sun-java 6-jdk
Hadoop	Hadoop-0.20.2

5.2 THE SYSTEM PERFORMANCE TEST AND ANALYSIS

The detection method, that the client uploads and downloads file from the server, is taken into use to detect the transmission rate change by different file size. Three different kinds of transportation methods, which are namely, unencrypted, AES encryption and DES encryption, are adopted to process the same file size. The file size is chosen from 0.2GB to 1.2GB. And these sizes of the file are more universal. The time consumption of uploading file is as shown in figure 3 and 4. In Figure 3, the vertical axis represents the time cost.

From the figure, experiment results show that the way of unencrypted is overall faster than the way of encryption. Using encryption to upload files, the speed of encryption algorithm AES is slightly faster than that of DES. In terms of user experience, the more cost time cased by encryption operation, is still acceptable. The situation of downloading is almost the same with uploading files.

In the process of transferring files, the system CPU occupancy rate should be also taken into consideration. According to the upload and download operation, the system CPU occupancy rate is detected by the program NMON which is used for monitoring. And the uploading operation CPU occupancy rate is as shown in figure 5.

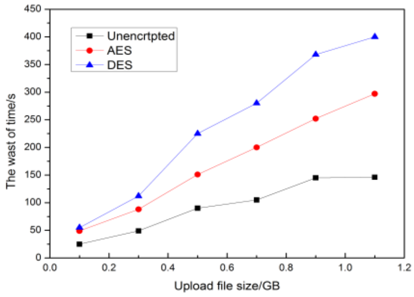

FIGURE 3 The time consumption of uploading file

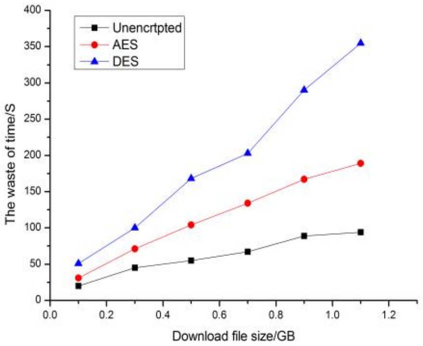

FIGURE 4 The time consumption of downloading file

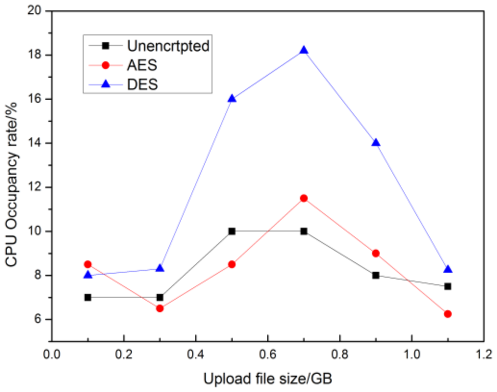

FIGURE 5 The uploading operation CPU occupancy rate of client

5 Conclusions

In this paper, Aiming at the client access, identity authentication scheme is designed for the data node. In this method, with abandoning the third-party trusted institutions, the principle of the algorithm of class IBE is used to simplify the key generation and reduce the key management complexity of storage. At the same time, due to the identity authentication scheme, the loss of data is avoided, which is caused by the malicious or counterfeit access to data node of client without through the name node.

References

[1] Kondo D, Javadi B, Malecot P, Cappello F, Anderson D 2009 Cost-Benefit Analysis of Cloud Computing versus Desktop Grids *Proc. IEEE Int. Symp. on Parallel&Distributed Processing (IPDPS09), IEEE Comp. Soc. Washington* 1-12

[2] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal 2009 Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities, Keynote Paper *Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications (HPCC 2008, IEEE CS Press, Los Alamitos, CA, USA), pp. 25-27, Dalian, China, Sept. 2009*

[3] Qinlu He, Zhanhuai Li, Xiao Zhang 2010 Study on Cloud Storage System Based on Distributed Storage Systems *Computational and Information Sciences (ICCIS), 2010 International Conference* **1**(2) 1332-5 17-19 Dec, 2010

[4] Weiwei Lin, Chen Liang, Liu Bo 2011 A Hadoop-based Efficient Economic Cloud Storage System *Third Pacific-Asia Conference on Circuits, Communications and System* 1-4 Wuhan, July 2011

[5] Weiwei Lin, Liu Bo 2012 Hadoop Data Load Balancing Method Based on Dynamic Bandwidth Allocation *Journal of South China University of Technology (Natural Science Edition)* **40**(9) 42-7

[6] Arfeen M A, Pawlikowski K, Willig A 2011 A Framework for Resource Allocation Strategies in Cloud Computing Environment *Computer Software and Applications Conference Workshops (COMPSACW), IEEE 35th Annual* 261-6

[7] Guo Lizheng, Zhao Shuguang, Shen Shigen, et al. 2012 Task Scheduling Optimization in Cloud Computing Based on Heuristic Algorithm *Journal of Networks* **7**(3) 547-53

[8] Kennedy J, Spears W 1998 Matching Algorithms to Problems: an Experimental Test of the Particle Swarm and Some Genetic Algorithms on the Multimodal Problem Generator *Proc IEEE*

International Conference on Evolutionary Computation. Piscataway, NJ: IEEE Service Center 78-83

[9] Huberman B A, Adamic I A 2004 Information Dynamics in the Networked World *Lect. Notes Phys.* **650** 371-98

[10] Qin I, Yu J R, Chang I 2009 Keyword search in databases: the power of RDBMS *SIGMOD Conference* **1** 681-94

Authors

Liu Zhilong, 1981-01-02, JiAn City, JiangXi Province

Current position, grades: lecturer
University studies: Computer science
Scientific interest: Database technology
Experience: From 2002 to 2014, working in XinYu University