Research on the Access Control and Communications Security Research Based on Hadoop Big Data Processing

Yanhua Hu[1, a]

[1]Department of Electrical and Computer Engineering, Lushan College of Guangxi University of Science and Technology, Liuzhou 545616, China

[a]huyanhua0220@163.com

Keywords: Hadoop, Big Data Processing, Access Control, Communications Security

Abstract. Hadoop is an open cloud computing platform. Cloud computing is a new computing model. It derives from distributed computing, grid computing etc, integrating available valid resources and providing computing resources, storage resource for user as service. The object faced by cloud computing could be enterprise, government, personal and so on, but Cloud computing is transparent to user who could not know the detail operation of the step in it, which can give user a feeling of high performance computing and large storage space and it begin to blend in human life. However, Cloud security issues have been blocking the development of cloud computing. To the majority of users, security is more important than computing power.

Introduction

The nature of cloud computing is a software concept, which aims to dig through a variety of software technology, integrate, manage hardware resources, after the large-scale computer resources unity, the computing resources as a service to the user. Cloud computing virtualization technology to achieve mainly through virtual technology was first put forward by Oxford University Strachey in 1959 and its goal is to improve the efficiency of the machine. Virtual technology is divided into a stand-alone virtualization and multi-machine virtual. Stand-alone virtualization through virtual technology to a machine as the use of multiple machines, their representatives have VMWare. Multi-machine virtualization goal is to unite as many machines a machine used by the control center, Hadoop is to represent its related technologies [1].

Cloud computing can integrate, tap various resources to form a powerful computing system, provide users with high computing performance, experience massive storage space, thus becoming the focus of current research and trends. But it has been accompanied by security issues, the development of cloud computing puzzle, how to ensure that users can safely use cloud computing and cloud computing to promote faster and more extensive use of a direct relationship. Some organizations or institutions because of the current cloud computing technology cannot provide effective protection of data security choose to abandon the use of cloud computing technology. Information age, information data representing all, have superior access to data, data processing capability is important, but the protection of privacy, security, important data is also important.

Hadoop is a cloud platform, has developed rapidly in recent years, cloud computing has become a hot research subject researchers. Hadoop open source performance so that more researchers understand cloud computing cloud computing, cloud computing depth, to promote the development of cloud computing to make contributions. But Hadoop is also facing security problems which faced by cloud computing, so improve the security of Hadoop data processing platform is particularly significant [2].

The Security Study on Cloud Computing

Cloud computing security problems both from traditional virus attacks, data theft, identity forgery and other issues, there are also large-scale network from cloud computing, virtualization, and new problems caused by the distributed features. Ref. The three basic cloud computing service software as a service (SaaS), platform as a service (PaaS), infrastructure as a service security threats and challenges (laaS) face are described. Cloud security issues analysis, cloud data security weak mainly reflected in the following aspects: (1) transfer security: data during transmission can be intercepted, but the transmission of data does not preclude the use of strong encryption Safeguard. (2) access control: access control permissions weak, user data stored in the cloud, and did not set a strong data access, users lose the absolute right to monitor the data. (3) Data Storage: user data after uploading the cloud may be distributed storage, users do not know the specific location of data storage, data confidentiality and non-confidential data was not classified storage and other factors may cause leakage of data.

The Architecture of Hadoop

Hadoop is an Apache's development distributed computing platform. In a distributed file system HDFS and MapReduce Hadoop core of the system to provide users with low-level details transparent distributed infrastructure. HDFS high fault tolerance, high scalability and other advantages allows users to deploy Hadoop in inexpensive hardware, form a distributed system. MapReduce distributed programming model allows users to develop parallel applications in distributed without knowing the underlying details, so users can take advantage of Hadoop easily organize computer resources to build their own distributed computing platform and can take advantage of cluster computing and storage the ability to complete the processing of massive data. Hadoop as a platform for enterprise-wide data processing infrastructure, with multiple subprojects as an aid to complete the appropriate function, such as HBase, Hive, Pig and the like. As in the previous section HDFS and Hadoop MapReduce constitute the two core. HDFS realization of the underlying support for distributed storage, and implement MapReduce distributed parallel processing tasks to program support and both support the Hadoop entire architecture.

The Overview on Hadoop Security

Although there are many Hadoop-related security researches but did not get certification authority, more authoritative security mechanism is only 2009 sets to Hadoop's Kerberos authentication protocol [3].

Hadoop0.20 version is less secure, it will separate the data stored on separate nodes, but no data is very strong protection, can easily be compromised. For solutions to this problem is to prevent unauthorized access to HDFS, all users must be authorized to access HDFS, including running MapReduce jobs, task and submitted by Oozie Job like. And users should also be authorized to authenticate the server, otherwise through forgery and counterfeiting servers may steal confidential information. Because the cloud node is transparent to the user, the user when uploading data to send confidential information to the server process data can be intercepted, and the service is counterfeit original server, taking the user's information confidential even get vouchers taking information on the server. Therefore, to prevent the occurrence of such incidents, users should also be authenticated to the server, unless the basis of communication is established in a closed, absolutely safe environment, or to the possibility of the existence of leaks.

The Secure Communication Design Based on Mixed Verification of Identity and Data

The System Structure. Structure associated with the way the application is also related to the specific implementation, the scope of application of different structure and role play are different. The system uses C / S structure enables the implementation process a little easier, because C / S model development technology is more mature, C / S system architecture has its own client, you can more easily control the client's data. But the drawbacks of C / S model is thought of Client Service must install the appropriate client software, and if the client software is required to run the client terminal Assembly in resources there are certain configuration requirements. Further C / S mode, the PC and embedded devices (IPod, etc.) require client software may not be the same, but also between different operating system client software will be different, such as Linux and

Windows systems Client software is different, use the IOS smart phone and use android operating system for smart phones and other smart phones are not the same for their clients [4].

In order to eliminate differences in the client, we propose preclude taking B / S design patterns. Customers can connect to the Internet anywhere you want data access can log on through a browser, you can get the desired service after obtaining appropriate permissions. Access through this way, the client can eliminate differences, more convenient and efficient access to required services. At the same time this design pattern is also coincide with the development trend of cloud computing. Try to reduce the client's request, the transfer of services to the cloud, to the client to provide a virtual terminal, access to services results.

The System Performance. In order to maximize the performance of the entire communication it is proposed to encrypt the data at the time according to the data of different security classification levels of data encryption. Data encryption to ensure that the data stored in the cloud will not be peeping, encryption can enhance data security, but also increase the burden on the system, not all user data needs to be encrypted, non-confidential file does not have encryption, such as the common document, files. So the client needs to determine before initiating a request to upload files and encryption type, in order to reduce unnecessary performance overhead. Therefore, this paper is encrypted according to the security classification level, the higher the higher the level of encryption security classification level, low level of encryption security classification level may be relatively low.

The System Security. Our aim is to provide data security in the transmission, storage process, while the server to ensure that the data has not been tampered with, replacing the need to digitally sign data, to ensure that the data received is indeed correspond to the user's data, the client and authentication servers are required to provide in order to ensure communication object claimed identity is indeed its own identity, it has not been forged. Here we were on authentication, data encryption, digital signature described in detail.

The Authentication. Authentication is used to verify the identity of the communication objects, whose purpose is to prevent unauthorized users from legitimate user faked data on the user or server theft, eavesdropping, tampering and other illegal behavior. It requires the user to provide authentication for cloud data protection is of great significance.

In order to do its utmost to protect the cloud data security purposes, the proposed scheme requires a combination of third-party authentication center AC. Security Center is a recognized third party for customers and cloud service center common trust. The third party requires a key center, which maintains an account of the domain Account Databases, the database records for each server and client domain name, user name, and the master key is called Master Key. Master Key non-dense steel itself, but the key through a hash function to obtain a hash value, since the hash function is not reversible, so there equivalence between the hash value and the key itself, but by such a calculation It can be well protected secret of steel is not compromised.

The Data Encryption. In order to have a better applicability, reduce the terminal equipment requirements, the program architecture designed for B / S architecture, so when a user to upload data, the signature of the user data encryption and data in the browser side. From a security point of view, to encrypt data to more effectively protect the data will not be easily peeping, to protect the security of the data. But the performance of the above analysis, the data encryption process will consume a certain amount of resources, more encryption and upload longer and higher the level of encryption resources consumed. Therefore, if the trade-off between these two factors requires a tradeoff. Therefore, according to the pros and cons of the program between the two, select the encryption scheme proposed, namely, encryption security classification according to the degree of the user's files [5].

System Security Analysis. (1) Authentication. Communication in an unsafe environment, providing authentication, which can effectively prevent unauthorized users pseudo causes legitimate users access privileges. (2) Transport key. Transport keys take the session key is transmitted in encrypted form. The session key is a temporary key with a life cycle, is more secure than the average long-term key. When the session ends, a session key also ends the life cycle. (3) End to end encryption. Data encryption on the client, upload the data encryption to protect the data will not be peeping in transmission, disclosure, while data is uploaded to the Hadoop after its contents cannot be seen directly, but also to protect the safety of its storage. (4) Data signature. User uploaded data is sampled on the sample data for data signature, the server changes the data examined, the owner can identify the received data is not claimed user. (5) Data to authenticate the user for data sampling signature, one can prove the identity of the owner of the data, on the other hand can detect whether the data has been tampered with in the process of transmission.

Conclusions

The security of cloud computing has been a sensitive, hot issue. Just because cloud computing security problems have not been solved, many potential users of cloud computing choose to abandon the use of cloud computing. As a cloud platform, Hadoop has got applications by a lot of enterprise and has been studied by many researchers, but Hadoop faced with the same security issues with the foregoing. Therefore, this article researched the access control and communication security for Hadoop big data processing, hoping to provide a reference to solve cloud computing data security issues.

References

[1] Kang Chen: Journal of Software, Vol. 5 (2009) No 20, p.25-26

[2] Yongwei Wu: Computer Research and Development, Vol. 12 (2011) No 27, p.46-55

[3] Shan Wang: Journal of Computers, Vol. 10 (2011) No 34, p.42-52

[4] Yu Zhang: Information Security and Communications Privacy, Vol. 11 (2005) No33, p.49-54

[5] Wei Xue: Computer Research and Development, Vol. 11 (2003) No40, p.29-33