A View on Data Security System for Cloud on Hadoop Framework

Karthik D
Dept of ISE
Acharya institute of Technology, Bangalore

Manjunath T N
Dept of ISE
Acharya institute of Technology, Bangalore

Srinivas K
Dept of ISE
Acharya institute of Technology, Bangalore

ABSTRACT
To solve the current data Security problem for cloud disk in distributed network, for example transmission, storage security problems, access control and data verification, a network cloud disk security storage system based on Hadoop is proposed. Based on the different secret level of client data, it provides selective encryption method which provides full deliberation to the subsequent security issues, such as the security of the client data broadcast in the network, client data no verification , the client data privacy might be leaked. Mutual with symmetric encryption algorithm and uniqueness of client data authentication of RSA, the performance of Hadoop, the distributed network cloud data security storage disk can supply protected, efficient, constant effect.

Key words: MapReduce, HDFS, Hadoop Cluster, Cloud Computing

1. INTRODUCTION
Cloud computing is an expansion of grid computing and distributed computing, which is a software concept in reality, it works through mixture of technologies such as software technologies, integration, management, and the use of various hardware sources. Cloud computing is realized mainly through the virtual technology [10]. The virtual technology can be separated into single virtualization and multiple virtualization; the single virtualization apply the virtual technology on a machine to work on numerous machines working commonly such as VMware, while multiple machine virtualization ties the machines through the control center and makes them work like single machine. Hadoop is the representative of the related technology, as shown in Fig 1. Obviously, its storage structure is distributed. The distributed storage system stores the data in similar devices which are independent of one another

Cloud storage is an important component of cloud computing, which is used to accomplish the target of placing data in the cloud. The network cloud disk, which is popular in modern years, is one of the ways to recognize the target. Logged users can store data in the cloud in a browser without any external storage media, and the user can get the data wherever they are by ordinary computers, mobile phones, laptop, iPad, etc .Cloud disk can make the storage easy, quick, and suitable, but the clients are not allowed to upload private data for security reasons. Security problems limit cloud disk from making improvement and expansion. The security problems of cloud disk are not only the conventional problems but also innovative problems in cloud computing. Described the security hazards and challenges of the cloud computing with its three basic patterns (Saas, Paas, Iaas)[10]. Cloud disk's weak security mostly occurs in the following aspects: Transmission security, Access control, Data verification, Data storage.

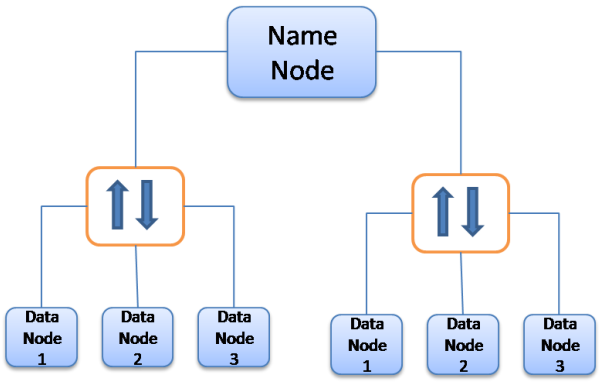

Fig 1. Hadoop Cluster

To solve the existing security problems, we suggest a cloud disk storage based on Hadoop, the program sketch lessons from Kerberos' authorization process. We make use of the classic algorithms such as AES, RSA to understand encryption, and authentication, and we also verify the time to inspect if it can inclusive the encryption and transmission in an acceptable period.

Hadoop is an Apache open source project which consists of HDFS, MapRedcue, HDFS common and HDFS YARN. It's main parts are HDFS and MapReduce. MapReduce aims at paralleling and dealing with tasks on a large scale, which would make the MapReduce scheduler become particularly important. HDFS is an open source project of Google distributed file system (GFS). It has a high fault tolerance and certain data access control. We mainly use Hadoop's HDFS (Hadoop Distributed File System) for cloud storage. As Hadoop also lacks safety measures, Kerberos was integrated into the Hadoop in 2009 by yahoo. The users have to obtain access certification from the third party center for key issues before access to Hadoop cluster. first, and it greatly reduced the risk of user's data leakages caused by identification. A lot of researchers proposed many different methods to increase the security of cloud storage proposed the use of the HDFS to build a private enterprise cloud, which combine the Hadoop's fault tolerance and suitable for big data attribute.

In SSL secure connection and secure virtual machine monitor are evaluated. encrypted the cloud data using attribute encryption (ABE) scheme, using the property as a public key to encrypt the data before it is uploaded, this limits the data access user to have K attributes to decrypt the data, in which the K is the number of threshold to decrypt the data. This scheme ensured the safety of data storage, and at the same time, the server has no need to keep a public key for each

user, the users' attributes are used to be the user's public key, but they could decrypt the data completely when different users hold their attributes together and get all attributes, identifies the users with image processing methods, such as face recognition, fingerprint recognition, etc. And made the transmission, encryption transmission and storage of the key files by Symmetric encryption and asymmetric encryption features complementary. But all these schemes mentioned just encrypt the data or identify the user from one perspective for a single demand, but no comprehensive data protection is taken. Information security is currently one of the most important issues in information systems. Security criteria most commonly used are confidentiality assurance that information is shared only among approved persons or organizations, integrity, availability and traceability

The problem becomes even more difficult if a user wants to export a document from the information system to work offline, for example, or to broadcast it to other persons outside the company. Drawing on the object-oriented approaches, we chose to encapsulate in the document itself some security components to achieve autonomic data management architecture for Enterprise Digital Rights Management (EDRM).Hadoop is an Apache open source project which consists of HDFS, Map Reduce, HBase, Hive and Zookeeper and other projects. Hadoop has two primarily parts: Hadoop distributed file system (HDFS) and Map Reduce programming model .HDFS is an open source version of the Google GFS implementation, as a highly fault-tolerant distributed file system, which provides high throughput data access, suitable for mass storage (PB-class) of large files (usually more than 64M).

Hadoop framework adopts Google cloud computing core computing model MapReduce which is a distributed computing technology and Simplified distributed programming model MapReduce computing model is divided into two parts, Map and Reduce. MapReduce model split a calculation job into a number of Map, and then assign to different nodes to compute. Programming input data to each Map job, after this step will input the intermediate data to Reduce job. Reduce job is to aggregate the Map intermediate data together and output. From the high-level abstract view. The HBase master is responsible for boot-strapping a virgin install, for assigning regions to registered region servers, and for recovering region server failures. The master node is lightly loaded. The region servers carry zero or more regions and field client read/write request.

2. LITERATURE SURVEY

2.1 Research on framework for urban Railway massive data based on cloud computing platform

In the current urban railway information platform, mainly consists of front-end data acquisition unit, backbone network and the backend server, the data collected by front-end acquisition equipment is transmitted to the backend server by backbone network [7].

There are many problems in this paper, one of the most important are:

1) Along with the further development of wireless technology and information industry, the traditional urban railway information platform has been difficult to adapt to the trend of the future. For example: the urban railway should be diversified mode of payment in the future, however, the traditional urban railway information platform obviously difficult to deal with this kind of demand. Another example, the current urban railway scheduling is mainly depending on the train operation diagram completed before operation. In the future train scheduling system should be able to real-time scheduling based on real-time traffic state [7]. Again, the future broadcasting system of urban railway should cover the whole city, all passenger can always check the running status of the train, station and line traffic conditions for planning their travel plans in advance, to save travel time, optimize the social resources.

2) The tradition platforms are expensive, not easy to maintain, difficult upgrade, poor reliability and so on. At present, due to the safety requirement of urban railway, all devices services are special, for example: there are system maintenance workstation, device maintenance workstation, permission maintenance workstation, running monitor workstation, statistics workstation, financial workstation, plan workstation, ticket workstation, auditing workstation, decision-making workstation *etc.* in operation management center, each service runs only one set of software. This design wastes many equipment resources, and it's necessary for the urban railway safety [7].

However, it's not enough for the safety of urban railway, each service must has its own backup server which instead of the primary service when it is fail. Similarly, the diversity of devices has brought the complex maintainable, urban railway company has to set up multiple jobs for the device maintenance. Secondly, it's difficult to upgrade. A vast amount of equipment, the complex data relationships, lead to when one of service upgrades others service appearing different types of problems, and it's necessary to suspend some devices during the progress of upgrade. Likewise, if one service is damaged, other devices will be affected seriously, hard to fast recovery. Finally, it created information island problems that artificially separate the data in order to enhance the system robustness. The data between the service nodes is not shared, not general, not synchronized and mutual interference.

2.2 Policy-based security framework for Storage and computation on enterprise Data in the cloud

Sourya Joyee De and Asim K. Pal [8] proposed a decentralized, active and developing policy-based security framework for enterprise data and computation outsourcing to the cloud and also it allows an group to retain control over its data and computations while being both well-organized and flexible. Here, they attempt to address this issue widely. They give emphasis to the storage and computation needs to be addressed discretely, integrated manner. They elaborate on the set of policies which they call the safe data policies consisting of storage security policies, upload security policies and computation security policies to guide the organization in finding out the correct security level for all combination of data security requirement and apparent adversarial actions on storage and computation nodes (VMs) of the CSP. The framework develops a people centric highly evolving and dynamic organizational view of the outsourcing operation of the enterprise data vis-avis the cloud. Further, the framework is based on the principle of danger minimization while optimizing efficiency and flexibility. They discuss the building blocks such as how the user at the individual level as well as the enterprise at the association level express their data

security requirements and CSP trustworthiness based on which we arrived at the security policies. Lastly, They suggest a possible association level implementation of the beyond security framework.

2.2.1 Secure Storage and Computation in Cloud

A number of recent works in cloud computing focus on storage and computation security. They propose a system architecture allowing organization-broad integration of un trusted community storage cloud. The architecture confirms confidentiality, availability and integrity while require only a minimum level of trust on the cloud. It uses the Information Dispersal Algorithms (IDA) to make sure availability, and by combining symmetric encryption with IDA, achieves high confidentiality. Integrity is make sure by using AES-CMAC operation mode for encryption which creates a MAC for all data fragment and enables replacement in the case of some integrity violation. [6] They presents a similar, advanced architecture where the end-devices within an organization are considered to be within a private Secure Cloud or π-Cloud, restricted by the π-Box that acts as an mediator between the π-Cloud and the external cloud. π-Box perform all security operations for data storage and sharing such as information dispersal, encryption, checksum etc. Data is initial dispersed using an IDA, encrypted, signed and then the shares are spread to multiple clouds. When the user inside the association needs to access data, the shares are obtained from the multiple clouds and the data is recreated if sufficient shares could be withdrawn.[6] proposes a cryptographic cloud storage service consisting of the following parts:

1) A data processor that processes data before being sent to the cloud;
2) A data verifier that verifies whether data stored in the cloud has been interfered with;
3) A token generator that generates token to allow the CSP to recover customer data segments and
4) A credential generator that implements access control policy by issuing documentations to various parties in the system.

It allows integrity, confidentiality in addition to secure data ensures. They also suggest to make use of searchable encryption to allow privacy and retrieval of data based on keywords and attribute based encryption to allow implementation of credentials and proof of storage to confirm integrity.[7] suggests a general-purpose protocol for secure computing any role in the cloud without revealing any information about the input or output by using multiple VMs. The usage of principles of secure multi-party computation (SMC) ensures that if at least a single VM is honest, no data is revealed. They derived from the literature of secure multi-party computation. [5] has proposed the double Cloud architecture for securely outsourcing data and arbitrary computations to the cloud. It consists of the usage of two types of clouds, the trusted cloud (such as a private cloud) which performs all security critical actions such as encryption, decryption etc

2.3 High Performance distributed Framework for standalone data analysis Packages for Hadoop-based cloud

The Hadoop Map Reduce is the programming model of designing the scalable distributed computing applications, that gives the developers to achieve automatic parallelization. However, mainly the complex manufacturing systems are tough and limiting to migrate to private clouds, due to the platform mismatched and complexity of system reconstruction. For growing the efficiency of manufacturing systems with minimum attempts on modifying source codes, a high-performance framework is designed and called *Multi-users-based Cloud-Adaptor Framework (MC-Framework)[9]*, which gives the easy interface to users for fairly executing requested tasks worked with traditional standalone data analysis packages in MapReduce-based private cloud environments. Moreover, this framework focuses on multiuser workloads, but the default Hadoop scheduling scheme, i.e., FIFO, would increase delay under multiuser scenarios. Hence, a new scheduling mechanism, called *Job-Sharing Scheduling*, is designed to explore and fairly share the jobs to machines in the private cloud. Then, we prototype an experimental virtual-metrology module of a manufacturing system as a case study to verify and analysis the proposed MC-Framework[9]. The results of our experiment indicate that our proposed framework enormously improved the time performance compared with the original package.

3. SYSTEM ANALYSIS

3.1 Cloud disk system

Cloud disk's weak security mainly occurs in the following aspects [10]:

3.1.1 Transmission security:

Data in communication process may be intercept, but the data communication is not working with the strong encryption security measures.

3.1.2 Access control:

Access control power is weak, the client data stored in the clouds without setting access power, the client lost absolute right to monitor.

3.1.3 Data storage:

Client upload data after the clouds, it is likely to be distributive stored, Client's do not know the exact position where the data is stored. And the private data and non-private data stored are not classified, which may cause the outflow of data.

3.1.4 Data verification:

The cloud makes no confirmation and examination on the data uploaded. It can't guarantee that the uploaded data is corresponding to the right client's data or the unique data from the client.

Disadvantages:

- Access control power is weak
- Outflow of data
- No confirmation and examination on the data Uploaded

3.2 Hadoop distributed file system (HDFS)

HDFS is an open source project of Google distributed file system (GPS). It has a high fault tolerance and secures data access control. The HDFS cluster consists of a Name node and a master server administers the file system namespace

and regulates admission to files.

Advantages:

- Reduces the load on the server.
- It provides Security, Stability, Efficient and successful storage

4. SYSTEM ARCHITECTURE

The common design for cloud disk structure comprising of the different end devices, such as smart phone, laptop, tablet PC, etc. The client could use a browser to login and right to use the data, which lowers the requirement by the terminal. The operation is easy, suitable and fast. Inspecting from its physical structure, the system is divided into three components, the client, server and cloud group [10], as shown in Fig2

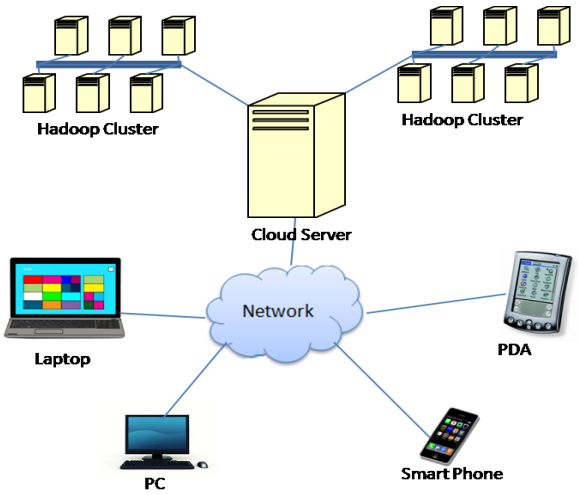

Fig 2: Cloud Structure

According to the software component, the system can be separated into client component, server call component, secret key production and distribution component, data encryption component, data signature component, the data transmission component, data authentication component, and data storage component, in which the client component includes the data encryption component, data transmission component ,data signature component while the server call component includes the secret key production and distribution component, data authentication component and data storage component[10].

5. CONCLUSION

Aiming at the existing popular cloud disc security weakness, we put advance a security encryption methods based on Hadoop which assures the data transmission and storage security and satisfy the server executes digital signature for client data at the same time[10]. It is a distributed encryption system that could reduce the load on the server, and lastly achieve security, stability, efficient and successful storage.

6. REFERENCES

[1] XU Guang-hui.Deploying and researching Hadoop in virtual machines[C]//Automation and Logistics (ICAL), 2012 IEEE International Conference on Zhengzhou, 2012: 395-399.

[2] LIUK. The Security Analysis on Otway-Rees Protocol Based on BAN Logic[C].Computational and Information Sciences (ICCIS), 20 Chongqing, 2012: 341-344

[3] MUNIER M. Self-Protecting Documents for Cloud Storage Security[C].Trust, Security and Privacy in Computing and Community, Liverpool, 2012: 1231-1238

[4] SHAIKH FB. Security threats in cloud computing[C].Internet Technology and Secured Transactions (ICIT, Abu Dhabi,2011: 214-219.

[5] V. Winkler, "Securing the Cloud Computer: Security Techniques and Tactics," Elsevier Inc., ISBN: 978-1-59749-592-9, 2011.

[6] Jianfeng Zhao "Research on Framework for Urban Railway Massive Data Based on Cloud Computing Platform", 2014 IEEE.

[7] Sourya Joyee De and Asim K. Pal "A Policy-based Security Framework for Storage and Computation on Enterprise Data in the Cloud" 2014, 47th Hawaii International Conference on System Science.

[8] Chao-Chun Chen, Nguyen Huu Tinh Giang, Tzu-Chao sLin "MC Framework: High-performance Distributed Framework for Standalone Data Analysis Packages over Hadoop-based Cloud" 2013 IEEE International Conference on Granular Computing(GrC).

[9] A. Huang Jing, B. LI Renfa, C. Tang Zhuo "The Research of the data security for cloud disk based on the Hadoop Framework" 2013 Fourth International conference on Intelligent control and information processing(ICICIP),Beijing, China.