Research on Kerberos Technology Based on Hadoop Cluster Security

Peng Shen[1, a], Xiaoming Ding[2,b] and Wenjun Ren[3,c]

[1] Colleage of Computer and Information Science in Southwest University, ChongQing,China

[2] Colleage of Computer and Information Science in Southwest University, ChongQing,China

[3] Colleage of Computer and Information Science in Southwest University, ChongQing,China

[a]pengshen.cs@qq.com, [b]2366268114@qq.com, [c]1453155741@qq.com

Keywords: Hadoop cluster, Kerberos, Security.

Abstract. This article is based on the ecological security of the Hadoop cluster. Combined with the practical application of Kerberos security authentication system in Hadoop cluster, we propose a multi node HA (High Available) configuration in Kerberos, realize Master-Master-Slave design mode and increase redundant nodes in storage. Based on the characteristics of dual Master, we set timed tasks, execute scripts on time, carry out data dump of Master nodes, prevent single Master failure, and avoid the failure of Kerberos authentication system. We use Slave nodes to carry out one way backup storage from dual Master to Slave. After double Master Kerberos database has been destroyed, it can recover Master nodes with the fastest speed. This improves the security and resilience of access to Hadoop cluster components and authentication architecture using database data.

1 Introduction

With the popularity of computers and the rapid development of IT technology, digital information is a major trend in the future. Large-scale data and information processing will soon become an unavoidable problem. The individual storage space is increasing with the demand of the society, and the total volume of data in China and the world is growing more and more every year. The storage size of the data center is rapidly moving towards the PB level or even the EB level. The arrival of these massive data means that the traditional data processing platform and processing model can not meet the needs of the reality. The platform and technology suitable for storing and computing these massive data appear. This is the basic framework of distributed system. The core design of Hadoop is HDFS and MapReduce. HDFS provides storage for massive data, and MapReduce provides computing for massive data. The emergence of Hadoop also means the arrival of the big data age.

In 2017, a "worm" type of extortion virus software WannaCry was spread by illegal elements. It caused the outbreak of the extortion virus in the world, covering more than 150 countries and hundreds of thousands of users, causing a loss of nearly $100 billion. Users of some Windows operating systems in China are infected. Users of campus network bear the brunt and suffer serious injuries. A large number of experimental data files and graduate design documents are encrypted and locked. When the application and database files of some large enterprises are encrypted, they can not work properly. Therefore, in the era of large data, the security of data should be put in the first place. Without all kinds of safety measures to escort data, the data will be destroyed and leaked. No matter it is a disaster for individuals or enterprises, the loss will be immeasurable. The focus of this article is the ecological security problem of Hadoop cluster. The lack of security verification mechanism can cause illegal users to easily invade the cluster. Illegal users can maliciously access components in a cluster and submit malicious jobs, disguised as other users to tamper with permissions and intercept or tamper with data on HDFS. Therefore, the Hadoop cluster needs a secure authentication mechanism to enhance the security of the entire cluster, and the Kerberos system meets this requirement.

2 Kerberos cluster authentication framework

Kerberos is a network authentication protocol which is designed to provide a powerful authentication service for client / server applications through a key system. The service can provide a powerful user authentication, and through validation, it can guarantee the real identity of the user. For each session, the security of all subsequent transactions in the session can be automatically protected by only one self validation to the service. In the Hadoop cluster, the role of Kerberos is the most important. The Hadoop Cluster Administrator wishes to hire the tenant of the cluster (hire the cluster user), and only use the corresponding components and related functions of its application, instead of using other components and related functions beyond their application scope. At this time, a authentication system is needed to authenticate and identify the identity of the landfall tenant so that it can obtain the appropriate use of the authority. In order to ensure the rational use of the cluster, avoid the misuse of unrelated tenants, or intentionally destroy the use of components that are not applied. At the same time, the benefits of doing this can also strengthen the management of the cluster, and allocate the relevant resources according to the needs.

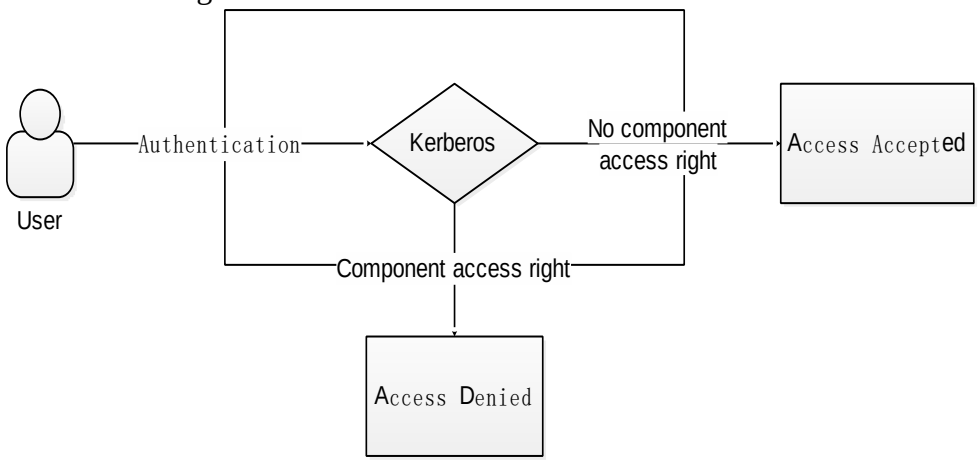

Fig. 1 The framework for tenants accessing the cluster through Kerberos

In a Hadoop cluster, the Kerberos system provides a method of transparent validation for the principal. In a cluster, principal can refer to a tenant or a related component of a cluster. The name of the component is regarded as the principle information in general,which means that the principle is the super administrator of the component.Taking Hbase as an example, if using keytab of Hbase components super administrator to authenticate, you can create a namespace, and to have access to Hbase privileges authorized tenants. The framework for tenants accessing the cluster through Kerberos is shown in Fig. 1.

3 Kerberos authentication principle

The user gets the Ticket T_{tgs} to visit TGS from the AS, then gets the Ticket T_s to visit S from TGS, and then gets the service from S. Kerberos authentication principles is shown in Fig. 2.

① C -> AS: C makes a requests to AS to get a ticket T_{tgs} (TGT + TGS Session Key)for access to TGS;

②AS-> C: KRB_AS_REP, AS goes to its own account database to find out whether there is the user's domain account, after authentication, responding to the request of C,and giving T_{tgs};

③ C-> TGS: KRB_TGS_REQ, C presents T_{tgs} to TGS, requesting access to S's ticket T_s;

④ TGS-> C: KRB_TGS_REP, after the TGS judgment, in response to the request of C, given T_s;

⑤ C-> S: KRB_AP_REQ, C presents T_s to S, request the corresponding service;

⑥ S-> C: KRB_AP_REP, S responds to C's request and provides the corresponding service.

In the Hadoop cluster, this section is transparent.The tenant only needs to obtain the key of the ticket principal from the specified key table,so as to authenticate the identity and get the service.

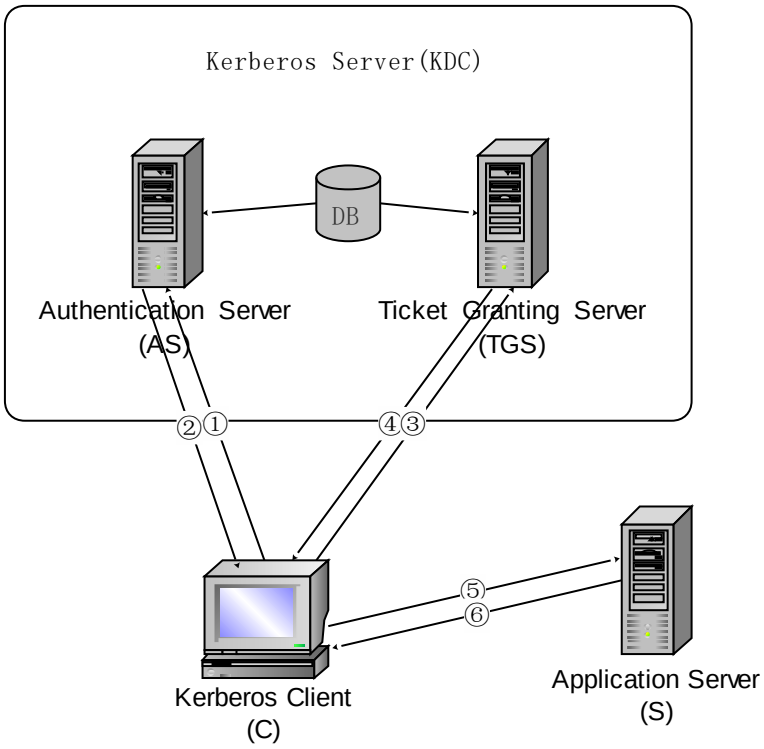

Fig. 2 Kerberos authentication principles

4 Master-Master-Slave model of Kerberos

This article combines the Kerberos service deployment in the Hadoop cluster and uses the Master-Master-Slave model to improve the security and stability of Kerberos in the cluster. The Master-Master-Slave model is shown in Fig. 3.

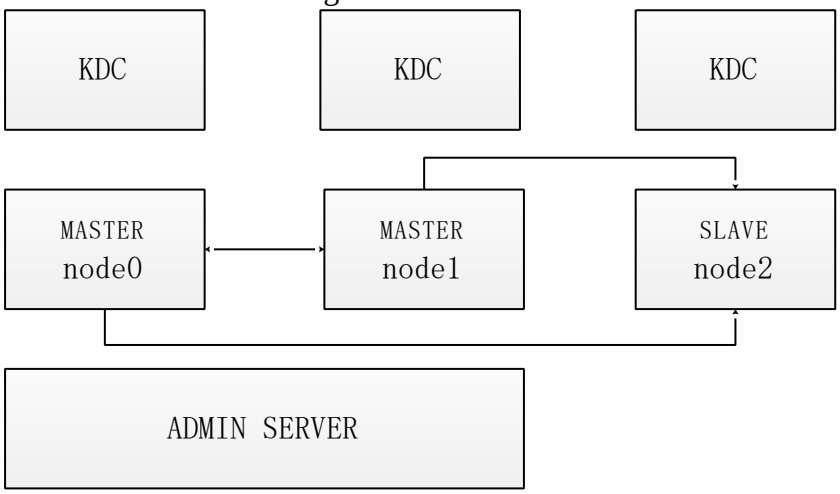

Fig. 3 The Master-Master-Slave model

This article proposes to set up two KDC servers on two more secure and stable nodes, and set a double admin server in the configuration file. In order, the system will default to the former as a master KDC server, and the latter is a slave KDC server. When the master KDC server fails, the slave KDC server detects the situation and begins to replace the master KDC server. The master and slave KDC server can all handle database management requests, and both can generate credentials. However, in order to synchronize the changes, the two needs to communicate with each other to

change information, so the master and slave KDC are set to allow the other to propagate the change information.

In order to prevent the loss of information from the KDC node, or a series of unexpected factors such as the destruction of Kerberos database because of the misoperation of the administrator account. Set timed tasks, execute scripts on time, carry out data dump of Master nodes, and use Slave nodes to store Master to Slave one-way, so as to achieve the fastest and best recovery of Master nodes with lost information. It is worth noting that, the configuration files krb5.conf, kadm5.acl, kdc.conf of nodes should be ensured to be the same. At the same time, When using kprop to store and propagate information in Kerberos database, it is necessary to ensure that all nodes' principal information verified by Kerberos is lowercase, otherwise it will lead to propagation error.

5 Master-Master-Slave model building

Experimental environment.

Server type: Red Hat Enterprise Linux Server release 6.8 (Santiago)

Server node is shown in Table 1:

Table 1 Server node

IP address	IP mapping
192.168.5.106	node0
192.168.5.107	node1
192.168.5.108	node2
192.168.5.109	node3
192.168.5.24	node4
192.168.5.47	node5
192.168.5.28	node6
192.168.5.111	node7

Kerberos client: krb5-workstation-1.10.3-57.el6.x86_64

Kerberos server: krb5-server-1.10.3-57.el6.x86_64

Configuration files: krb5.conf, kadm5.acl, kdc.conf

Experimental deployment. The RHE system needs to reconfigure the YUM source, otherwise there will be a downloading error. The Kerberos client is deployed on all nodes, and the Kerberos server is deployed on the node0, node1, and node2 nodes. The Master node is set to node0 and node1, and node2 is used as a Slave node, and the rest of the nodes are client nodes. Modify the three major configuration files under node0, and then synchronize to all nodes. The kpropd.acl file is created under node0, node1, and node2, which is the authentication subject information that allows information to propagate.

Experimental test. This model is designed to enhance the security of the cluster and the security and ease of recovery of the Kerberos system.It mainly involves the authentication of principal information, the information dump between master-master, the information dump between master-slave and the transmission of slave information to master nodes for data recovery. Based on the functional points involved, a test case is created to cover to detect whether or not it can reach the desired purpose and effect.

① Both the two master nodes and the slave nodes create principal and export their KeyTab files. These KeyTab files are transmitted to the specified directories of each node respectively, and the authentication tests of the users are carried out respectively. Verify that the authentication of the principal information of each node is normal by cross validation.

② Open kprop services, and transmit information between master and master, master, and salve for forward dump. Verify that the dump propagation is normal.

③ The Kerberos database in a master is destroyed, and then the data of the second step dump to the slave node is safely transferred to the master node for recovery. Detect the recovery of Kerberos.

The test results are consistent with expectations,which shows that compared with no Kerberos building, master-slave model of Kerberos or master-master model in Hadoop cluster, regardless of the

safety of the entire cluster, or Kerberos their own safety, the advantages of Master-Master-Slave model are relatively more obvious.

6 summary

This paper describes the theory and technology of Kerberos authentication based on Hadoop cluster in detail, introduces its authentication framework and principle in cluster, and puts forward the Master-Master-Slave model of Kerberos in practical application, and builds a cluster environment to prove it. The experimental results fully show that Kerberos's Master-Master-Slave model improves Kerberos's security and resilience under the premise of ensuring cluster's security authentication, and supplements and develops the original Kerberos authentication.

References

[1] D Hu, D Chen, Y Zhang, S Pei, Research on Hadoop Identity Authentication Based on Improved Kerberos Protocol, *International Journal of Security & Its Applications*, vol.9, pp. 429-438, 2015.

[2] Q Zhou, D Wu, C Tang, C Rong, STSHC: secure and trusted scheme for Hadoop cluster, *International Journal of High Performance Systems Architecture*, vol.5, pp.63-69, 2014.

[3] YL Wang, J Wang, An Access Control Policy of Big Data Storage Platform Based on Kerberos and HDFS, *Computer Engineering & Software*, vol. 37, pp. 67-70, 2016.

[4] Cheng `Xiaorong, Q Feng, C Dong, M Zhang, Research and Realization of Authentication Technique Based on OTP and Kerberos, *High Performance Computing and Grid in Asia Pacific Region*, International Conference on (2005), pp: 409-416, 2005.

[5] H Qiu, Y Quan, Research and Design of Kerberos-based Single Sign-on System, *Computer Applications*, vol.23,pp.142-144,2003.

[6] SM Bellovin, M Merritt, Limitations of the Kerberos authentication system, *Acm Sigcomm Computer Communication Review*, vol.20 (5), pp.119-132, 1990.

[7] ZZ Wang, Y Wang, Research and Design of Campus Network Unified Identity Authentication System Based on Kerberos, *Advanced Materials Research*, vols.546-547, pp.1086-1089, 2012.

[8] P Yang, H Ning, Security Analysis and Strategy Research of Kerberos Protocol, *Computer Engineering*, vol.41, pp.144-148, 2015.

[9] HK Abdalrazzq, MS Ibrahim, OA Dawood, Secure Internet Voting System based on Public Key Kerberos, *International Journal of Computer Science Issues*,vol.9,pp.428-434,2012.

[10] RR Parmar, S Roy, D Bhattacharyya, SK Bandyopadhyay, TH Kim, Large-Scale Encryption in the Hadoop Environment: Challenges and Solutions, *IEEE Access*, vol.5, pp.7156-7163, 2017.

[11] Z Dou， I Khalil， A Khreishah， A Al-Fuqaha, Robust Insider Attacks Countermeasure for Hadoop: Design and Implementation, *IEEE Systems Journal*, vol.99, pp.1-12,2017.

[12] F Bach, HK Cakmak, H Maass, U Kuehnapfel, Power Grid Time Series Data Analysis with Pig on a Hadoop Cluster Compared to Multi Core Systems, *Euromicro International Conference on Parallel* , pp.208-212,2013.

[13] Z Ren, J Wan, W Shi, X Xu, M Zhou, Workload Analysis, Implications, and Optimization on a Production Hadoop Cluster: A Case Study on Taobao, *IEEE Transactions on Services Computing*, vol.7,pp.307-321,2014.

[14] SM Bellovin, M Merritt, Limitations of the Kerberos authentication system, *Acm Sigcomm Computer Communication Review*, vol.20 (5), pp.119-132, 1990 .

[15] G Bella, E Riccobene, Formal Analysis of the Kerberos Authentication System, *Journal of Universal Computer Science*, vol.3 (12), pp.1337—1381, 1997.

[16] E El-Emam, M Koutb, H Kelash, OS Faragallah, An Authentication Protocol Based on Kerberos 5, *International Journal of Network Security*, vol.12 (3), pp.159-170, 2011.

[17] L Ma, Y Zhu, An enhanced Kerberos protocol based on one-time password, *Icic Express Letters*, vol.8 (9), pp.2497-2502, 2014.

[18] A Boldyreva, V Kumar, Provable-security analysis of authenticated encryption in Kerberos, *Iet Information Security*, vol.5 (4), pp.207-219, 2011.

[19] JA &Lt, Informational: Kerberos GeneralString to be Interpreted as ASCII Only, *European Polymer Journal*, vol.70, pp.203-214, 2015.

[20] Y Ling, YS Jin, ZH Zhou, One Three-Point Authentication Protocol Base on Kerberos Protocol, *Bmc Gastroenterology* , vol.13 (1), pp.287-301, 2013.