The Research on Big Data Security Architecture Based on Hadoop

Miao Zhuang[1, a]

[1]Xi'an International University, Modern Educational Technology Center, Shaanxi,xian,70077,china

[a]miaozhuang@xaiu.edu.cn

Keywords: Big Data, Security Architecture, Hadoop

Abstract. This paper will discuss the security of big data from two aspects: the first is the architecture, we will list some differences in handling and storing information in big data systems, and discuss how they can affect the security of data and databases. The second is the operation and maintenance management, the operational safety issues with big data platform-dependent will be described in detail, and analyze the challenges facing the large data and management system, especially when the system itself lacks the internal security mechanism. Finally this paper gives the technical advice, outlines strategies and tactics of large data security to ensure the safety of these data warehouse.

Introduction

Big Data security is a deceptively simple and very difficult issue, there is no mature solution. Big data more widely, thousands of enterprises engaged in related work with big data, development of new technologies that allow companies to manage and analyze big data, meanwhile, has become increasingly large databases. When people find themselves are quite dependent on "big data", they begin to aware of the value of big data and protect it [1].

Big data has the following characteristics: PB level or above data processing; distributed redundant data storage; parallel processing tasks; providing MapReduce or similar data processing functions; fast data insertion; centralized management and business processes; relatively inexpensive; hardware-independent; easy access; scalability.

The Introduction of Hadoop

"Commercial Bank Personal Financial Services Interim Measures," September 24, 2005 published in the Bank of personal financial services and professional services to clearly define the activities of asset management, investment advisory, financial planning and financial analysis for commercial banks to individuals, mainly for the commercial banks to use more of the unified management of investment funds, investors with financial products with the bank signed the agreement made some gains, and bear the risk. Commercial bank financial products is a similar collection of brokerage asset management plan Capital Trust scheme and securities investment fund financial products, commercial banks in the market potential target customers on the basis of systems analysis for a series of specific objectives customer base design with concurrency and selling a capital investment management solutions. According to different operating mechanism and management model, business financial products attributable to the commercial bank financing of integrated financial services business [2].

Hadoop is calculated based on the open-source Java implementation of cloud-based framework developed by the Apache Foundation. It appears so that the user can design their own cloud computing platform and develop cloud computing applications. Currently, Hadoop has become the industry standard platform and academia cloud computing applications and research. Hadoop framework (combined HDF, YARN, Common, etc.) is the product of big data, which provides all the necessary features. Most of Big Data Hadoop system uses one or more components, or extend some basic functions.

The Simple DB of Amazon can meet the requirements of large data and despite the Hadoop architecture is different. Google's BigTable architecture is very similar to Hadoop. Paper will focus

on the Hadoop framework and its derivative products NoSQL environments, Hadoop framework is actually a stack, you can freely add and remove, for example, Sqoop and Hive is an alternative to mutual data access services component. Because they are not subject to relational database structures or relationships Query Analyzer constraints, you can choose a different query engine based on the stored data type. You can use tools such as Scribe HDFS extended logging capabilities, the entire stack can be configured and expanded as needed. This modular approach provides great flexibility but it makes security more difficult, because it does not consider itself safe mechanism for each component [3].

The Safety of Big Data Security Analysis Architecture

The paper needs to build an architecture model to ensure the safety of large data cluster. Compared with traditional databases, data warehouses, and massively parallel processing environments, large data security challenges facing different, many of the software early in the design did not consider the security mechanisms embedded. Hadoop file system builds highly distributed, redundant and resilient data storage pools. Distributed File System provides a number of essential features, and the massively parallel computing possible. But the layers of the stack integration and communication between the data node and the client / resource management agencies and it will introduce new problems.

The problems of architecture are as follows [4]:

The distributed nodes. "Handling computing resources is cheaper than transporting data" and which is critical for large data. In a highly distributed data cluster, it is difficult to verify the consistency of security between heterogeneous platforms.

The data fragmentation. Data Big Data is flowing within a cluster and there are multiple copies of data moving between different nodes to ensure redundancy and resiliency. In other words, data fragmentation is shared server data. This makes it difficult to move or copy data promptly and accurately locate where data is stored, or to know how many data backup is available. This is the traditional model of centralized data security is breached, the centralized data security model, a single copy of the data processing are protected by different mechanisms.

The data access. Role-based access control model for most databases are very common. Property relational databases including roles, groups, templates, "safety label so as to restrict user access to a subset of the available data authorization. Most big data environment provides a Schema-level access control, but there is no finer granularity in the Big Data environment, learn safety labels and other advanced properties is possible, but it requires application designers to integrate these capabilities into applications and data storage to go.

The communication of inter node. Hadoop and the vast majority of communications between components is unsafe, they use RPC TCP / IP over. And TLS and SSL security mechanisms are not embedded in large data communications.

The interaction of clients. Clients need to interact with resource managers and nodes. Although you can create a gateway node is used to load the data, but the client can direct resource managers a single data node communication with the parties or. Malicious customers can send malicious code or a link to any service. This is very efficient in terms of communication, but it is difficult to ensure that data from the client node attack or clients from data node attacks. Worse self-organization, distributed node architecture for gateways, firewalls, monitoring tools use is not appropriate, which requires the design of a distributed control point adaptation peer mesh cluster and now do not have.

Lack of security mechanisms. Finally, and perhaps most importantly, big data stack design itself does not consider security mechanisms. No facility can be used to protect the data warehouse, applications or Hadoop core functions. All large data equipment is built on Web services model, it is difficult to deal with common threats Web network. As we all know, most of the API vulnerable.

The Operation Management Security

Big Data security includes security data and infrastructure, so to protect the application (database) and management data in order to protect the management of information. If desired, an attacker can directly access the data, bypassing the database management system, unless there is a direct path to access the information, otherwise the attacker will look for weaknesses or attack database applications [5].

In addition to the structural safety of the System of Hadoop and similar platforms, but also lacks many common security controls, like the IT management team maintains other data management systems, which includes security configuration management and access control, as well as built-in auditing and security gateway service capabilities. Of course, there may be selected NoSQL component will provide one or two security component, which could be a Web proxy, network data encryption, or full bi-directional authentication access management components. But usually it is a single point of security features, rather than a comprehensive set of tools. Worse, safety is always just an option, not enabled by default, even if the Hadoop Web console allows access without any form of authentication.

The following is a common threat data management system, and if the cluster administrator to set up or manage big data cluster, can be considered for security control from the following aspects:

Static data protection. The standard for the protection of static data is encrypted, preventing the establishment of an external application interface to try to access the data. Similar to traditional data management systems, worried man stealing files or read files directly from disk. Encryption cannot prevent users from accessing the file encryption key. Copy effectively replaces backup big data, but it does not mean that a malicious administrator or cloud service managers will not secretly create your own backup. One or two NoSQL component may provide static data encryption, but most will not provide. Most encryption products lack the scalability and transparency.

Management data access. Each node has at least one administrator can have full access to their data. Compared with encryption, it needs a border or a facility to separate duties between different administrators. Relational database platform is similar, but the lack of big data platform built-in tools. If you do not want to directly access the data or the processed data, by combining access control, the role of classification, encryption technology or the like.

Configuration and patch management. In a cluster of servers, with the passage of time, the new node configuration and patch level may not be the same, if you use a different operating system platform, which will determine the level of patching bring difficulty. Node cluster can tolerate reused without data loss or service interruption, but restart could cause serious performance problems, depending on which nodes are affected, and how to configure a cluster.

The application and the node identity validation. Hadoop can use Kerberos to authenticate users, but malicious client can access the network, if the Kerberos credentials stolen or copied, can also use a virtual image is taken from a file or a snapshot of credentials. When embedded credentials in virtual and cloud environments is more need to worry, as it is more readily available client applications exact copy. The nodes strong authentication is a very important tool to prevent malicious access to your server cluster.

The auditing and logging. If you suspect someone destroy or visit your cluster, how to detect it? You need activity records. Scribe and LogStash; open source tools can be integrated into most big data environments, as some commercial products. So, just you need to find a compatible tool to install and integrate with other systems.

Monitor, filter and block. Built-in monitoring tools to find misuse or block malicious queries. If you care about security, Kerberos clients can authenticate access to MapReduce must pass the message digest authentication. Some monitoring tools can be used for large data environments, but most of the work at the API level.

The security of API. Big Data Cluster API need protection from injecting malicious code and attack commands, buffer overflow attacks, Web service attacks. Common security controls include

an integrated directory service, the OAuth token is mapped to the API service, filter requests, input validation and policy management across nodes. Some do not even API authentication mechanism.

The Technical Advice

Use Kerberos in nodes authentication. Kerberos can effectively verify communication between services, blocking malicious nodes in the cluster and applications. It can protect access to the Web console, making it difficult to be attacked management channel. Kerberos authentication node and the new application requires additional overhead, but if not establish a two-way trust, easy to integrate into Hadoop malicious applications or malicious nodes. Kerberos is one of the most effective security controls, and can be integrated into the infrastructure of Hadoop.

Use file-level encryption. If a malicious user or administrator tries to gain access to the data node direct access to files, the encryption protects data and stolen or copied file is unreadable disk image. File-level encryption to provide consistent protection and some products even can protect encryption memory. Equally important is the encryption should comply with the safety requirements of big data and Hadoop ---- This application is invoked is transparent.

Use the key management functions. If an attacker can get the encryption key to encrypt the file will fail. Many large data cluster administrator is convenient to consider the key is stored in a local disk drive, but this insecurity, because the key can be acquired platform administrator or attacker. Using Key Management Service to distribute keys and certificates, and set a different key for each group of applications and users. Most of the encryption depends on the key / certificate security.

Use the secure communication mechanism. Between nodes, between nodes is required with the application SSL / TLS components to achieve secure communication. Transfer large data set on the cluster to bring a small loss of performance, but the burden is shared among all nodes.

Conclusion

The article discusses the security in the environmental of "big data," and the embedded protection mechanism and vulnerability in the inspection system. Now it seems to that the realization of big data is highly dependent on low-cost clusters, safety considerations tend to lag behind. The results show that these deployments are largely unsafe, entirely dependent on the network and perimeter security support. Many clusters are deployed in virtual and cloud environments can use the management tool provided by vendors to address many security problems. But the security in architecture and operational level should be strengthened and develop data security technologies.

Acknowledgement

2015Scientific Research Program Funded by Shaanxi Provincial Education Department(Program NO:15JK2134)

References

[1] Z. Y. Pan, The information security and big data, J. Information Security. 24 (2011) 59-61.

[2] J. Bai, When big data meets information security, J. Information Security and Communications Privacy. 16(2013) 84-87.

[3] X.R.Ye, The mobile e-government system based on Android platform, J. Technology Review, 12(2010) 189-190.

[4] Z.G.Xiong, The research and implementation on cloud computing development, J, University of Electronic Science and Technology, 4(2012) 58-63.

[5] W.Cui, The research and practice on data integration technology in distributed environment, J. University of Electronic Science and Technology, 9(2011) 43-50.