The analysis of big Data application mode and security risk

ZENG Tai-sheng[123]

[1] Faculty of Mathematics and Computer Science, Quanzhou Normal University, Fujian, Quanzhou 362000, China

[2] Fujian Provincial Key Laboratory of Data Intensive Computing, Fujian, Quanzhou 362000, China

[3] Key Laboratory of Intelligent Computing and Information Processing, Fujian Province University, Fujian, Quanzhou 362000, China

Keywords: big data; security analysis; Hadoop

Abstract. In the current social background, Existing facilities and tools already can not meet the needs of big data in expanding and analysis techniques. Today's data storage and analysis work is achieved under cloud conditions and Hadoop were set up. Under the conditions for cloud computing, cloud computing applications who have remote data files were not authorized to read its contents, which results in unauthorized manipulation, and it will produce a lot of security risks for large data. In this paper, according to the cloud of different modes, Hadoop different stages of the operation, subject to the threat of non- confidence and security to steal big data generated to analyze a variety of privacy, with threat model as an example, it explores ways to address security threats.

Preface

At present, when it comes to communications, data generation rate in computing, which continue to grow rapidly, while the data processing capacity of traditional data processing is more limited, so for these much-needed data, you must choose the research and development of new technologies to fight. Traditional data processing modes slowly cannot meet the increasingly large number, increasingly complex structure, expanding a wide range of data, traditional analytic tools is facing enormous challenges. Usually people only take the form of adding compute nodes and systems data processing functions can be realized. Therefore, a highly efficient and reliable data-processing pattern--cloud pattern has emerged. Through cloud computing platform and calculation module of Hadoop can efficiently handle large streams of data. Due to the characteristics of the model itself, which results in relatively weak security, and brings a certain amount of risk.

Big Data application mode

Cloud computing is a way of computing over the Internet, which has just started the calculation. Parallel computing, distributed computing, grid computing, the three calculation methods have been widely used before. Cloud computing is based on the calculation of the three original for the development and extending cross. In addition, Cloud computing is a fusion of virtual network technology.

The use of cloud computing, which makes people no longer have to pay big bucks to purchase or maintain part of the infrastructure of computer, for the company, even more significant. Because cloud computing can share professional storage and computing devices to store personal data needs to be stored. So, some specialized computer equipment configuration and maintenance, which are considered by the cloud computing service providers to carry out. Companies or individuals want to use this shared resource, they only need to follow certain rules of valuation to pay the rent.

With the development of the company's horizontal integration, outsourcing technology becomes widespread,which are more and more able to be accepted by the majority of companies. Cloud computing is undoubtedly an advantage of development opportunities, which can be a good opportunity for outsourcing. Because of the nature of cloud computing, it is obvious that it is the best choice for large data storage. Therefore, some companies should consider outsourcing companies by big data in the cloud to service providers, and improve the profitability of enterprises.

At present, with the development of cloud computing. Large data has already been recognized as the implementation plan. This programme is funded by Map in the Google cloud computing

Reduce the Google File System technologies implemented the open source Hadoop, Hadoop is essentially an accounting framework. This calculation framework takes Hadoop Distributed File System and Map Reduce technical as the center, which provides a distributed processing system capable of handling high throughput and large data. Hadoop is the most mainstream of the treatment program. There are two ways to run Hadoop, the first is to build a cluster to run the organization through their own strength; The second is to run up through lease other hardware. Among them, the second method is used more widely. There are mainly two leased hardware cloud services, one is public-private partnerships which can be used of Cloudera, one is called Elastic Map Reduce from Amazon's cloud services.

Thus, it can be seen that the large data is processed through the cloud and Hadoop cross, which has increasingly become a mainstream way of dealing with large data.

Big Data Security Risk Analysis

Any application of a wide range of technology or computing approach, there are safety considerations. Big Data has been able to get public recognition, an important reason is security. However, as the complexity of the network environment, data security has become an urgent need to overcome the difficulties.

Cloud computing is a way to calculate outsourcing. This also means that users lose the ability to directly control one's own resources. Because, once using cloud computing means that users already have the right to subcontract service providers. It is worth noting that, in the cloud computing, big data is not encrypted. Service providers have the right to manage the data. Therefore, there will be a case of theft of data service providers. In addition, the central platform for cloud computing will be subjected to malicious attacks from third parties. If it can successfully destroy its security and defense, then third parties can control illegal and to read the data. This will pose a great threat to the security of customer data.

Hadoop in the initial design, the safety factor is not considered as a category. However, after a subsequent version Hadoop1.0.0 version and Cloudera CDH3 in two versions, one can be called the Kerberos authentication mechanism is added. And join them together, and have always access control mechanism. This control mechanism is based on security considerations, which is based on the extension of the ACL technology. However, since these two safety mechanisms are not able to authenticate the application platform, which can only be carried out on the machine level security certification. Therefore, its security is also not be valued by people. This is mainly due to the Kerberos only by Clients-- i.e. The client, the KDC Key Distribution Center, and Server-- servers, mutual authentication between the three. The security strategy based on ACL is more complicated.

Security strategy based on the ACL, and it can only begin to run after the user has enabled ACL mechanism. After the user is enabled ACL, the need for nine attributes in the value of the hadoop-policy.xml property is set again. These properties require the administrator to configure them so that once changed, can not easily be gap. And because of its own security and defense mechanism is weak, changing its configuration is relatively easy. However, For different customer needs, which is provided in accordance with the need to make further changes to the customer's individual requirements. This nine properties for Hadoop resources can restrict access to areas, can also restrict Jobtracker and Tasktrackers or node communication of Datanode and Namenode. Moreover, its size therefore can not be protected by the fineness of the heart to meet customer standards to protect their privacy field. And, because of its security is not high, the operation is not convenient, which is not conducive to system maintenance features. The security mechanism is not perfect.

The risks of CSP and Uers on different application modes

In cloud computing, the application model of Hadoop is more than one. When it is built on a private cloud, with good closure . Some companies choose their own Hadoop applications, and user of the platform control to the company's employees, who set the blocking access permissions for foreign people. The overall risk between CSP and Uers are shown in figure one.

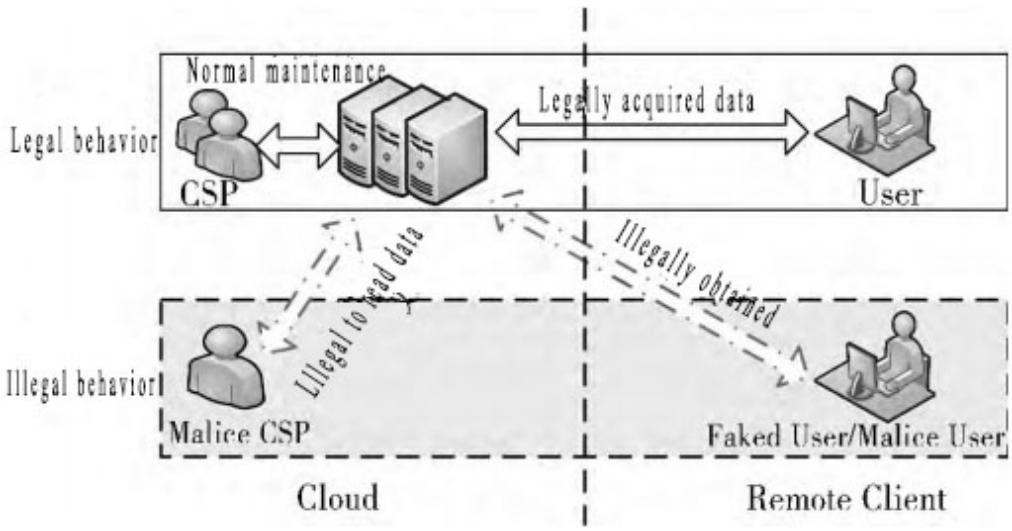

Figure one A risk map between CSP and User

Threat Model

It refers to the risk between CSP and User in last section, and then we analyze theft data between the two issues in detail. Second, as shown in Figure two, this figure includes four objects, which are those data storage, data adopters, data operators and Internet service providers. The picture is taken from the simulated threat scenarios information Airvat system, which is designed from the threat model.

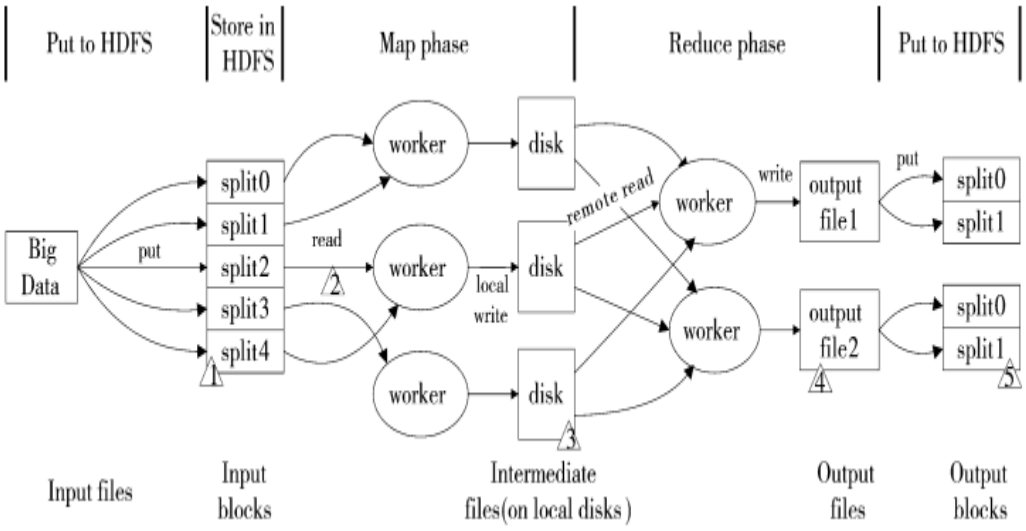

Figure two The analysis of data illegally obtaining position

In a Big network vendors Shop company, through years of operation, which accesses to large amounts of user information, and now to some extent assumptions. Store user information used is the date of the company records, orders, customers, it has a substantial customer transaction database. Now supposing that there is an intelligent robot Rresearcher, pay a remuneration to Bob,designed the system data exchange model is directed backward mining, then data changes on the system to Hadoop framework, followed by data analysis specially prepared Map Reduce code, wherein automatically divided in accordance with different roles. Data users and Internet service providers set up for Big shop, set the data storage for the customer, and then take the robot as a data analyst. Suppose a very special day, intelligent robots want to get the day of the transaction order system, so it can automatically write a Mapper code gets generated for each record, if today's date is C, after which the output of the key / value pair <K, order>K. K is a sequence of strings, reduce only summarize the relevant of each key k, and outputs the result, and related user information will be leaked out, such disclosure may result in Big Shop in a disadvantageous position in business,

also disclose transaction data, and suffered a credibility problem without customer consent.

Through an appeal to the actual threat model, we can sum up the threat which has several of the following characteristics:

(A) When there is malicious behavior of Internet service providers, in order to convenient for the user information data type, usually adopt methods to improve their system privileges, and through the middle of the copy system data is stored in one place, so that there are other effect in the future.

(B) In the offline state of the cloud computing platform, that is security mechanism fails, the hacker can take some direct means of obtaining the highest authority, so as to achieve the purpose of data tampering, and data falsification and other illegal activities.

(C) The identity of the person by using the data calculation, type some important information in the map, and analyze through Map Reduce code.

Security Policy

In the context of big data model, which is called data services security that is to protect files and file systems, and which is known as privacy protection that is to store the key content and delivery of the results of computer protection. At the time of safety issues for consideration, it should be done in the privacy of the behavior of these two angles and data come together. This article will elaborate about security policy:

First, both the CSP and the user identity as well as a platform from which all require a comprehensive verification. After the Kerberos mutual authentication, which confirms the identities of both CSP and guarantee its security, only if the request has been initiated to reach certification.

Second, after both the safety and reliability of the information are determined, it is necessary to verify the safety of the platform. It is based on authentication, trusted computer security, on the TPM to build cloud computing environment which can be trusted, then complete the platform environmental trust for authentication, link build trusted, trusted links to physical layer into the virtual layer, and so on to ensure the security of cloud computing.

Third, the data is used in the process, which also conducts real-time monitoring, such monitoring is illegal request CSP issued a denial, at the same time, the data can not normal computer operations and data output supervision. Using LSM settings, the corresponding control measures are set out, using hook function controls the kernel of legitimate data access, and strictly put an end to illegal data, so as to ensure the flow of data is a legitimate subject.

Conclusion

The primary content of this article is now provided for storing and analyzing big data, which is carried out under Hadoop framework for cloud computing implementation. Then it gives the book a different arrangement of cloud computing mode, Hadoop operations process run through the maze. Non trusted service on the data and privacy threats, which is combined with appropriate threat model example, so as to arrive at a credible platform, the main source of risk is accounted for by CSP and data. Saying this is because in the case of CSP, they will find the user does not change the permissions of user files. Accounting of data privacy when entering the paragraphs will get the relevant data, so data security is threatened. Finally, with a solution to this problem, I will continue the process of accumulation of knowledge, and continue to explore more innovative and effective safety program.

References

[1] Ma Yuan. Security mechanism study based on the Hadoop cloud computing platform[J]. Information Security and Communications Privacy. 2012 (06) : 78-79

[2] Zhou Tianyang, Zhu Junhu, Wang Qingxian. Based on the VMM Rootkit and testing technology research [J]. Computer Science. 2011 (12) : 256-257

[3] Zhang Xinyuan Li Baiyang. The concept of big data, technology and application [J]. Journal of innovation of science and technology. 2013 (9) : 127-128

[4] Wei Wei the banking industry cope with the challenge of Internet financial strategy under the background of a preliminary study, based on large data thinking [J]. Journal of fujian financial. 2014 (7) : 56-57

[5] Wang Yao The data processing method based on large data analysis [J]. Journal of digital technology and application. 2014 (6) : 123-124

[6] Guo Yajun. Big data era digital publishing service model innovation study [J]. Journal of economic study Tribune. 2014 (4) : 90-91

[7] Xie Tian, Tang Hao Time has come for a data explosion [J]. Journal of intelligence. (3): 2013-47 48