Research on the Cloud Storage Security in Big Data Era

Chen Kai [1, a], Lang Weimin [2, b], Zheng Ke [3, c], Ouyang Wenjing [4, d]

[1] The Eighth Department, PLA Academy of National Defense Information, Wuhan, 430010, China

[2] The Second Department, PLA Academy of National Defense Information, Wuhan, 430010, China

[3] The Third Department, PLA Academy of National Defense Information, Wuhan, 430010, China

[4] Department of Science Research, PLA Military Economics Academy, Wuhan, 430035, China

[a]email: chenkai817788@163.com, [b]email: wemlang@163.com,

[c]email: 237422762@qq.com, [d]email: 466071789@qq.com,

Keywords: Big Data; Big Data storage; Storage security; Cloud Computing

Abstract. In the Big Data era, users increasingly prefer to take clouds as major locations for data storage. With the presence of cloud computing, vast amounts of data bring huge social and economic benefits to all of us. But this also poses daunting challenges to the data storage security. The existing security measures can no longer meet the requirements of large data storage security. These threats will emerge as the new bottleneck of information security. As such, this paper classifies the major threats to cloud storage based on analysis on the defining features of Big Data. We also propose protective strategies for big data storage security from the prospective of key storage security technology, management and relevant policy-measures.

Introduction

With the rapid development of ICT, the world is witnessing a skyrocketing growth in information storage, whose measure has expanded from PB to ZB. According to statistics released by the International Data Corporation, the global data storage of 2011 has reached 1.8ZB. And it's expected to be doubled every 2 years. It is estimated that the global data will exceed 40ZB by 2020, suggesting every man on earth contributing 5,200 GB data [1]. Figure 1 showcases the trend of global data growth.

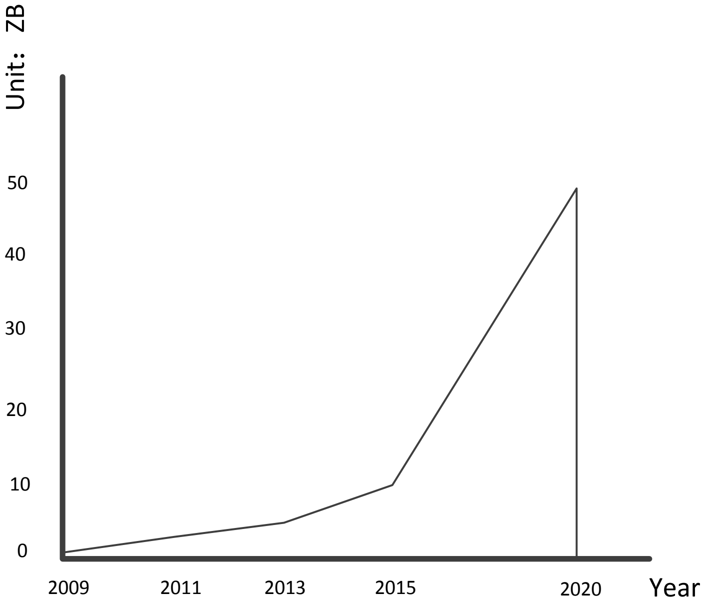

Figure 1: the Trend of Global Data Growth

Novel technologies such as the Internet of Things, Cloud Computing and Mobile Internet are

© 2015. The authors - Published by Atlantis Press

playing increasingly inseparable roles in our daily lives and changing the way people work and study. Developments in mobile and smart terminals, communications networking and software applications have greatly propelled the exponential growth of global data. Thanks to the upgrading and evolvement of Cloud Computing and other milestone innovations, Big Data technology was born.

However, Big Data might bring new challenges to data storage, transmission and processing as people across the world are now paying more attentions to information privacy. The security measures we employ nowadays can no longer satisfy the growing demand for a more secure Big Data environment. As such, it is imperative for us to develop and devise a highly efficient and urgent solution to Big Data storage security issues. In this paper, we approach the problem from the perspective of cloud storage security.

Features of Big Data

Currently, researches on features of Big Data focus on 3Vs and 4Vs. 3Vs represent the Volume, Velocity and Variety, whereas the extra V in 4Vs stands for Value [2], which is strongly supported by the International Data Corporation.

The vast Volume of data measured by PB will continue to exist and double every two years. Variety can be seen in the types and sources of data, e.g. the vast amount of structure, semi-structure and non-structure data and data sourced from the Internet of Things, Cloud Computing and Mobile Internet. Velocity means the rapid forwarding and processing of data. And the Value of data will diminish as the time lapses. The Obama administration defines Big Data as the oil in the future, suggesting its promising prospects. As displayed in figure 2, people pay different attentions to the 4Vs of Big Data.

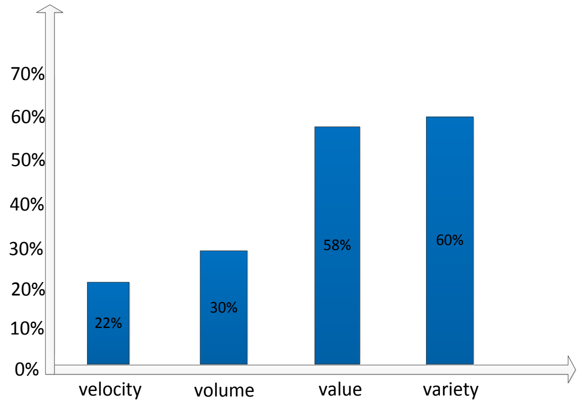

Figure 2: Attentions Given to the 4Vs of Big Data

Challenges to Cloud Storage Security

Big Data technology represents the strategic trend of development for future info technology. It will attract billions of money from numerous enterprises and institutions. Big Data will stand as a cornerstone for technological innovations. However, as the shining stars of ICT fields, Big Data and Cloud Computing come hand in hand. Cloud Computing specializes in data storage and distributed computation, whereas Big Data plays an important role in data filtering and mining. Cloud Computing lies as the foundation for Big Data technology which in turns fuel the growth of the former. The combination of the two is an urgent demand of our times [3]. However, with the surging increase of data, the security of cloud storage in the context of Big Data remains a problem to be reckoned with. The followings are main threats to cloud storage security:

a) Broken Access Control

Users illegally use network or data resources without consent. For example, attackers purposefully bypass system’s access control mechanism and illegally utilize network and data resources. They may gain privileges to access into sensitive information by ways of IP masquerading, identity attacking and unauthorized access [4].

b) Advanced Persistent Threat

APT is a kind of organized and secret attacks with pre-set purposes. It is carried out in varied ways with persistent efforts and can sometimes cause great mayhems to target hosts. In the future, APT is more likely to take Big Data as its primary goal. It might pose serious threats to information security as our current solutions to this problem are often limited. The process of APT attack is illustrated in figure 3 [5].

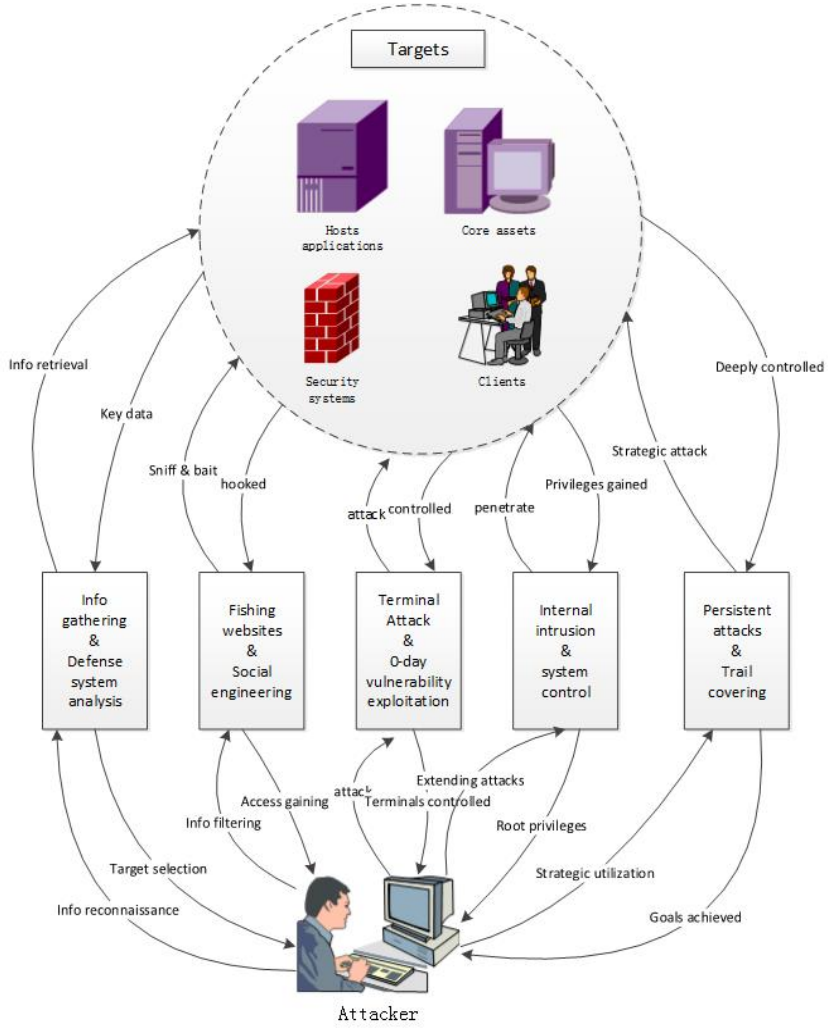

Figure 3: the Process of APT Attack

c) Privacy Disclosure

Big Data contains huge amount of identity and behavior information of numerous users. If those data are not properly managed or protected during their storage, transmission and process, then users' privacy might be undermined. And there are more serious threats to privacy disclosure as hackers might be able to predict users' behavior by analyzing their personal information, customs and habits. To put it in a simple way, your will be caught in a grave danger once your personal and private information is encroached upon. In light of the mismatch between data owners and data managers in the era of Big Data, traditional protective measures can no longer safeguard the security of Big Data. That is why privacy disclosure will emerge as an issue demanding urgent solutions.

d) Data Loss and Disruption

Cloud storage of the Big Data era could suffer from 3 types of data loss and disruption. The 1st one is data loss and disruption during transmission, e.g. attackers could steal sensitive information by tapping into the routes or links between two communicating hosts. The 2nd type is to distort information in the third party which holds the highest privilege. The 3rd attack aims at the network applications which are keys to cloud storage.

Protective Measures for Cloud Storage

In the Big Data era, cloud storage security faces many threats. Traditional storage security technology can't satisfy Big Data's demands any more. This is why many enterprises and institutions are still reluctant to embrace cloud computing in the near future. To address this problem, we propose the following measures:

a) Data Encryption

It's imperative for us to encrypt crucial data, especially in the era of cloud computing. Generally, there are two kinds of data encryption, namely, Symmetric Algorithm and Asymmetric Algorithm. Symmetric Algorithm uses the same key for data encryption and decryption, such as DES、AES、IDEA and RC4. On the other hand, Asymmetric Algorithm uses two different keys, namely, private key and public key. The public one is accessible to the public, whereas the private one is only known to a specific user. A piece of information is encrypted using the public key and then decrypted by the user who holds the private key, and vice versa. Normal Asymmetric Algorithm includes RSA and ElGamal. Since SA and AA have their own advantages and disadvantages, some scholars have introduced a combination of both algorithms, which in the case quite agrees with Big Data security.

With scientific and technologic advancement, clouds are becoming increasingly preferable places for users to store their data especially in the Big Data era. But this could often leads to potential data stealing and disruption due to the following reasons. Firstly, clouds providers might leak sensitive information as a result of servers malfunction. Secondly, once clouds are attacked and intruded, users' data might be exposed to distortion and disruption. In the connection, users must encrypt their data before uploading it to clouds and decrypt their data after downloading it from clouds. Only in this way can we ensure that sensitive data wouldn't be exposed even if they are leaked. The followings are introductions of main attribute-based encryption policies.

Attribute-based encryption policies are mainly divided into two different types [6-7]: Key-Policy Attributed Based Encryption (KP-ABE) and Ciphertext-Policy Attributed Based Encryption (CP-ABE).

KP-ABE uses tree structure to describe access policy. Au represents the leaf nodes set. The encrypted data is relevant to attribute set Ac. The data can be decrypted only when Ac match Au [8]. Figure 4 displays the mechanism of KP-ABE. We assume the attribute set of encrypted data as Ac: { Ma，Mb，Md}, and the attribute set of both users as user1: { Ma，Mb} and user2: { Ma，Mc}. Since only the Au of user 1 matches with Ac, user 2 has no access to information in the encrypted data.

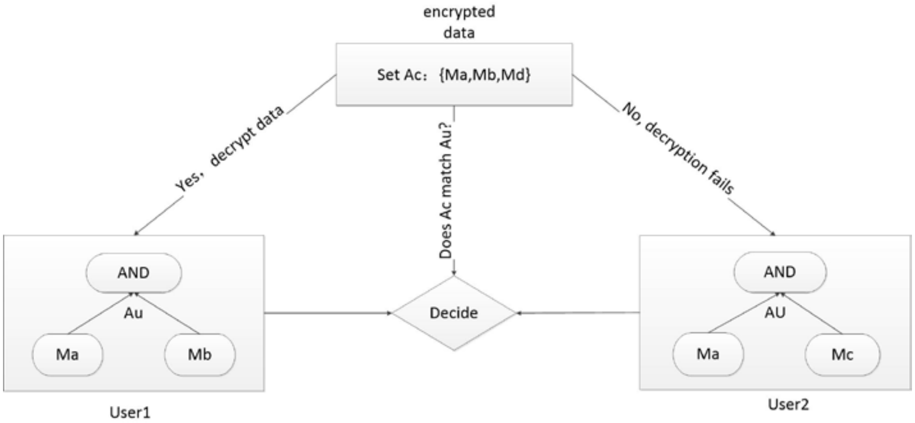

Figure 4: the Mechanism of KP-ABE

CP-ABE also uses tree structure to describe access policy Ac-cp, according to which the sender decides the control strategy. The key is relevant to the attribute set Au. The user can decrypt the data only when Au matches Ac-cp. The mechanism of CP-ABE is indicated in figure 5. We assume that user1 and user2 are relevant so that it's possible that both of them can decrypt the ciphered texts. But in the presence of Ac-cp, only user can decipher the text because the attribute Md of user2 does not match with Ma or Me.

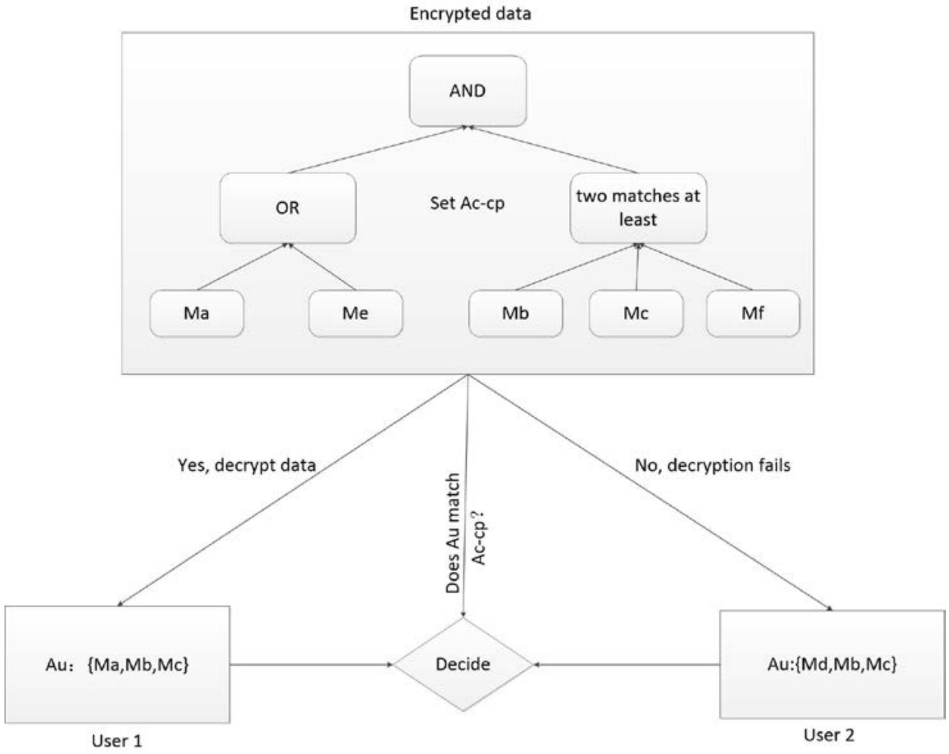

Figure 5: the Mechanism of CP-ABE

b) Identity Verification and Access Control

Identity verification refers to checking of users' identity in the computer network. Traditionally, identity verification is achieved based on passwords, letter of credits and biological features. However, these measures might be compromised if the passwords are stolen or the Credit Certificate is lost. Especially in the Big Data era, it's necessary for us to adopt a new identity verification policy for big data analysis. By acquiring and analyzing the behavior of a user, we will know the behavioral characteristics of the user and the device. Then we can identify the user by knowing his or her behavior. The advantage of big data analysis lies in that the attacker is impossible to mimic the behavioral features of a particular user. Hence, it can help us to fend off illegal intrusion as a result of passwords stealing or Credit Certificate loss.

Access control helps to limit access to crucial big data resources, improve the permission level, prevent illegal utilization of resources and safeguard information security. Regular access control measures include self-motivated access control, compulsory access control and role-based access control.

c) Data Backup and Recovery

Data storage security must be ensured by a robust backup and recovery mechanism so as to maintain data availability and integrity. In the case of data disruption or loss, a sound data backup and recovery system can help to ensure sustained data usage. Traditional data backup and recovery techniques involve Remote Backup, RAID, Data Mirroring and Data Snapshot. But these methods are designed for relatively small amount of data. For big data measured by PB, they are simply inapplicable.

Currently, Hadoop Distributed File System is widely used as a solution to big data backup and recovery. With advantages in copy conservation and backup, HDFS not only builds up data availability and reliability, but also contributes to network bandwidth efficiency. Still, it's necessary for us to build remote backup center. Things might be getting relatively simpler, owning to emerging cloud storage technology which is cost efficient, rapidly deployable, easily manageable and highly reliable. At application level, we can build a backup system by utilizing the virtualization technology and distributed parallel programming technology. At the data level, we could achieve the goal by taking advantage of cloud storage technology for its oblivious edges mentioned above.

d) Rules and Standards

As a novel and revolutionary technology, Big Data will be strongly supported by the government. It will greatly transform the country’s industrial mix and ways of management. As such, we must design and implement a complete set of rules and standards for its health development. Nowadays, countries across the global have laid down relevant laws concerning data security. Our government must also work to establish data management and security regulations so as to ensure the health and secure utilization of big data.

Conclusion

With the rapid development of Big Data era, every sector of the society is losing no time in introducing and spreading this technology for those great benefits it brings. But we must not be blind to the potential threats and challenges incurred along with it. Big Data technology can truly propel the development of our times only when we can strike a perfect balance between its applications and security.

References

[1]Zhihan Lv, Alex Tek, Franck Da Silva, Charly Empereur-Mot, Matthieu Chavent, and Marc Baaden. Game on, science-how video game technology may help biologists tackle visualization challenges. PloS one 8, no. 3 (2013): e57990.

[2]Yu Song, et al.. A Rapid and High Reliable Identify Program for Nighttime Pedestrians. Infrared Physics & Technology. 2015.

[3]Yishuang Geng, Jin Chen, Ruijun Fu, Guanqun Bao, Kaveh Pahlavan, Enlighten wearable physiological monitoring systems: On-body rf characteristics based human motion classification using a support vector machine, IEEE transactions on mobile computing, 1(1), 1-15, Apr. 2015

[4]Jie He, Yishuang Geng, Fei Liu, Cheng Xu, CC-KF: Enhanced TOA Performance in Multipath and NLOS Indoor Extreme Environment, IEEE Sensor Journal, 14(11), 3766-3774, Nov. 2014

[5]Shuang Zhou, Liang Mi, Hao Chen, Yishuang Geng, Building detection in Digital surface model, 2013 IEEE International Conference on Imaging Systems and Techniques (IST), Oct. 2012

[6]Xiaoming Li, Zhihan Lv, Baoyun Zhang, Ling Yin, Weixi Wang, Shengzhong Feng, Jinxing Hu. Traffic Management and Forecasting System Based on 3D GIS. 2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid). IEEE, 2015