Secured Storage of Big Data in Cloud

Melbin J Reena, A. Shajin Nargunam

Abstract--- Big Data refers to large volume of data and necessitates the usage of cloud for storage and processing. Cloud tenants data is not only stored in the cloud, but it is also shared among multiple users. The data stored in cloud must be well protected as it is prone to malicious attacks and hardware failures. Also, user's data on cloud contain sensitive information that must be protected and highly restricted from unauthorized access. Cloud deployment models such as public cloud, private cloud, and hybrid cloud can be used for storing data of cloud tenants. This paper proposes a secured storage approach for protecting data in cloud by partitioning big dataset into blocks containing user's sensitive data, insensitive data, and public data. Sensitive data is moved to private cloud and is well protected using proxy re encryption. Insensitive data is stored in public cloud and some data blocks are randomly encrypted. Also, the storage index information of insensitive data blocks on cloud is encrypted and shared among authorized users. Public data is also moved to public cloud and to protect it the storage path information is only encrypted and shared. The proposed approach shows better results with reduced computation overhead and improved security.

Keywords--- Big Data, Proxy Re-encryption, Storage Path, Cloud Storage, Sensitive Data.

I. INTRODUCTION

The emerging big-data paradigm, owing to its broader impact, has profoundly transformed our society and will continue to attract diverse attentions from both technological experts and the public in general. The big data paradigm recently has received considerable attention since it gives a great opportunity to mine knowledge from massive amounts of data [1]. In the cloud computing context, network-accessible resources are defined as services. The most beneficiary service models are Infrastructure as a service (IaaS), Platform as a service (PaaS), and Software as a Service (SaaS) [2]. Nowadays, there is an explosive growth in the volume, velocity, variety, veracity, and value of data produced over the internet. The distinguished characteristic of big data necessitates the usage of cloud computing for storage and processing. Cloud computing is being intensively referred to as one of the most influential innovations in information technology in recent years. With resource virtualization, cloud can deliver computing resources and services in a pay-as-you-go mode, which is widely preferred. Data security / privacy is one of the major concerns in the adoption of cloud computing. Compared to conventional systems, users will lose their direct control over their data. The challenge is how to ensure data confidentiality and integrity when storing such data in cloud.

Cloud deployment models such as private cloud, public cloud, and hybrid cloud can be used for processing big data. Private cloud is more secure as it is maintained by organization itself but there are some limitations also. The first limitation is scalability. On-premise private cloud deployments might not consider future growth, resulting in limited scalability. This is not surprising, as building highly scalable private clouds requires a large capital investment for procuring and installing computing and storage resources. However, the changing volume, velocity, and variety of data make it difficult to accurately plan private cloud capacity, and private clouds are often either under or over provisioned. To reduce capital investment, private clouds are always built with limited scalability.

Analytics is another possible limitation. Analytics models and software frameworks required to manage heterogeneous data might not be available in the private cloud because of higher operational costs. A third limitation is data sharing. Data must be shared with collaborators who don't have access to private clouds. Although private clouds are trustworthy, these limitations hamper the use of it for big data processing.

On the other hand, public clouds support the scalability and easy sharing of data. However, public cloud service providers and existing big data processing frameworks have no easy way of detecting or monitoring data leakage. Therefore, data auditing, data protection, and privacy preservation have emerged as salient areas for protecting data on cloud.

It is important to bring together the inherent features of public clouds and private clouds to build a trustworthy big data processing platform. In this paper, we proposed a mechanism for shifting data to hybrid infrastructure consisting of private and public clouds.

II. BIG DATA SECURITY IN CLOUD

Cloud computing has become the tool of choice for big data processing and analytics due to its reduced cost, broad network access, elasticity, resource pooling, and measured service. Cloud computing enables customers to store and analyze their data using shared computing resources. However, cloud computing comes with risks. The shared compute infrastructure introduces many security concerns not present in more traditional computing architectures. The cloud provider and tenants may be untrusted entities who try to tamper with data storage or computation. These concerns motivate the need for a novel framework for analyzing cloud computing security, as well as for the use of cryptographic tools to address cloud computing security goals. Protecting big data while it is in storage is a challenge for most of the organizations.

To maintain integrity and security of big data stored in cloud servers, sharing user's sensitive data in cloud searching need to be avoided.

Revised Manuscript Received on July 10, 2019.

Melbin J Reena, Research Scholar, Noorul Islam Centre for Higher Education, Kumaracoil, Tamil Nadu, India. (e-mail: mjreena82@gmail.com)

Dr.A. Shajin Nargunam, Director (Academics), Noorul Islam Centre for Higher Education Kumaracoil, Tamil Nadu, India. (e-mail: shajin@niuniv.com)

Security mechanisms must be devised to protect sensitive data from migration between cloud servers [5].Secure sensitive data sharing over cloud involves four primary safety factors. First, there are security issues when sensitive data are transmitted from a data owner's local server to a big data platform. Second, there can be sensitive data computing and storage security problems on the big data platform. Third, there are secure sensitive data use issues on the cloud platform. Fourth, there are issues involving secure data destruction. Issuing and renting sensitive data on a semi-trusted big data platform requires a data security mechanism. Building secure channels for a full data life cycle requires considerations of four aspects of safety problems: reliable submission, safe storage, riskless use, and secure destruction. A common and popular method of ensuring data submission security on a semi-trusted big data platform is to encrypt data[18].The disadvantage of encrypting data is that the user cannot share his encrypted data at a fine-grained level. When a data owner wants to share someone his information, the owner must know exactly the one he wants to share with. In many applications, the data owner wants to share information with several users [7]. To handle this complexity of encrypting data as a whole the proposed method encrypts only users sensitive data and also provides a mechanism of protecting data by distributing it over different cloud providers.

III. RELATED WORK

This section presents a review of existing works about protecting big data in cloud environment. [1] Proposes identity based generalized signcryption approach to obtain confidentiality and authenticity efficiently in big data. A collaborative intrusion detection mechanism is proposed in [2] for detecting cooperative attacks in cloud computing environment. A method based on PRE is proposed in [4]. A semi-trusted agent with a proxy key can re-encrypt ciphertext; however, the agent cannot obtain the corresponding plaintext or compute the decryption key of either party in the authorization process. A fully homomorphic encryption mechanism is also used to protect data. This mechanism permits a specific algebraic operation based on ciphertext that yields a still encrypted result. More specifically, retrieval and comparison of the encrypted data procedure correct results, but the data are not encrypted throughout the entire process. This scheme requires very substantial computation, and it is not always easy to implement with existing technology. [7] Proposes Key-policy attribute-based encryption with time-specified attributes to secure data in cloud. In this approach, attribute-based encryption is employed along with destructing sensitive data after use is also proposed. For verifying the integrity and correctness of data in cloud storage, privacy preserving public auditing scheme for shared data in cloud is proposed in [10]. The major advantage of this work is that it is not necessary to download data to verify its integrity. [11] Proposes public auditing of dynamic storage on cloud that can fully support authorized auditing and fine grained update requests. In [13], secure cloud computing based framework is proposed for big data information management of smart grid. This framework provides security solution based on identity-based encryption, signature, and proxy re-encryption. A framework for secure sensitive data sharing on a big data platform, including secure data delivery, storage, usage, and destruction on a semi-trusted big data sharing platform is proposed in [18]. [19] Proposes secure big data storage and sharing scheme for cloud tenants by protecting the mapping of the data elements to various storage providers. When compared with existing works related to protecting cloud storage, the proposed work suggests hybrid cloud framework for protecting cloud data.

IV. THE PROPOSED SCHEME

The cloud is increasingly being used to store and process big data. Cloud users need to split big datasets into smaller datasets and store them in different physical servers for reliability, privacy preserving or efficient processing purposes [2 4]. Cloud tenant's data stored on cloud comprises sensitive data, insensitive data, and public data. Sensitive data contain more valuable information that should be protected at any cost and so private cloud is preferred for storing it. Insensitive data is also important in the view of cloud user, but it can be shared at least with organization members and can be uploaded in public cloud. Public data is meant for sharing among potential users, who work as a group. This work proposes separate methods to well-protect user's data.

Sensitive data is protected by proxy re-encryption mechanism, so that it is not possible even for the proxy to view original data. To protect insensitive data, the storage path of data in cloud is encrypted along with encrypting some random data blocks. Encryption of only storage path information is simple as compared to encrypting entire data. Public data can be stored freely on public cloud and only the storage path information of data blocks is encrypted without encrypting data.

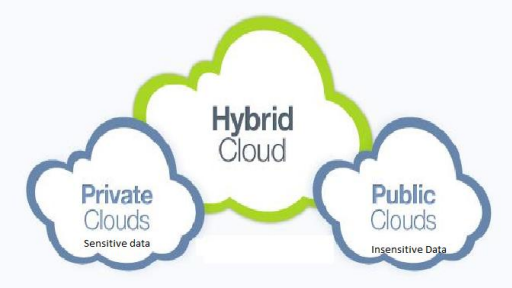

Fig. 1

4.1. Protecting sensitive data

The shared data in cloud contains user's personal data, health details, and financial data that need to be well protected. To protect this kind of sensitive data, proxy re-encryption(PRE) is used in this work.PRE transforms ciphertext meant for a user into a ciphertext of the same message meant for another user without obtaining decryption key or the actual message. Sensitive data is kept secret and is protected from unauthorized access by irrelevant users. Confidentiality is obtained by encrypting the contents of stored files.

The server operators can distribute encrypted files without having access to the plaintext files themselves.

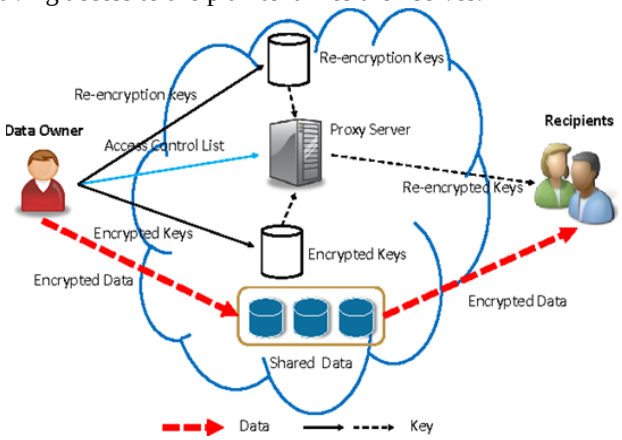

Fig. 2: Proxy Re-encryption

The cloud tenants protect their data by using proxy re-encryption in which a proxy server re-encrypts data without knowing the actual plaintext. Also, it generates re-encryption key for recipients on request for decrypting data.

Algorithm

PRE involves identity-based encryption, Re-encryption, and decryption for protecting data in cloud storage. Identity-based encryption includes **Setup, KeyGen, Enc, and Dec**.Re-encryption by proxy involves **ReKeyGen** and **ReEnc** for re-encryption key generation and re-encrypting the first level ciphertext and for sharing it with the intended user. The idea of PRE process is effective and simple in protecting cloud data. The encrypted data is moved to big data platform and the designated proxy generates re-encryption key, which is used to re-encrypt data. During re-encryption, even the proxy cannot obtain the clear text or decryption key. The re-encrypted cipher text can be decrypted by the authorized recipient after applying PRE services. For decrypting the re-encrypted text, the recipient can use its own private key. The step in implementing PRE algorithms is as follows:

Step 1: Setup(k): *the security parameter k is given as input. A primary security parameter msk is generated.*

Step 2:KeyGen(msk,id): *the identity of the user is obtained and verified by the key generation center. On providing legal identity, the private key(pk) and public key(p1k) are generated using primary security parameter msk.*

Step 3:Enc(pk,p1k,r1,d): *the message d is encrypted into first level ciphertext C1 using private key pk, and public key p1k under condition r1.*

Step 4:ReKeyGen(p1k,msk,r1): *the re-encryption key is generated for data owner by providing public key p1k, security parameter msk under condition r1. The data owner compute PRE key for another user, say j, using his identity. Thus PRE key* $rek_{idi\text{-}idj}$ *is provided to user j.*

Step 5:ReEnc(C1, $rek_{idi\text{-}idj}$***, msk,r1):*** *Re-encryption of first level ciphertet C1 is carried out using security parameter msk and* $rek_{idi\text{-}idj}$ *under r1 to obtain second level ciphertext C2 intended for user j.*

Step 6:Dec(C2,p1k,msk): *the intended receiver can receive the second level ciphertext C2 and can decrypt it to obtain clear text d using its own public key p1k.*

The proxy re-encryption is carried out on the big data platform to improve user convenience and to utilize the computing resources effectively.

4.2. Protecting Insensitive Data

Insensitive data of user uploaded on public cloud need to be protected as well. Sensitive data can be encrypted entirely to provide better security and access control. Big data sizes range from few terabytes to many petabytes and so it is not advisable to encrypt big data as a whole. This paper proposes a method in which the storage path information of data in cloud is only encrypted. The data blocks containing insensitive information can be stored across different cloud storage providers and the index file gets encrypted. When tenants demand for their data, data blocks can be collected from various locations and arranged by using partition identifier associated with every block. Application of trapdoor function for protecting the mapping details of data elements with different cloud service providers proved to be a simple approach than encrypting entire data.

Trapdoor function plays an important role in cryptography for protecting and sharing data among authorized users. The data partitions consisting insensitive data is represented as $block_x$ and $x\epsilon$ (1,n), where n is the total number of data partitions. The $block_x$ can be mapped to any of the cloud storage provider represented an sp_y and sp_y ϵ (1,m), where m is the cloud storage providers. To provide better security to insensitive data, some of the data blocks are encrypted randomly. The storage path information of data partitions is encrypted using trapdoor function and the result is represented as TF_{Tdval}(Storage Path), where TF stands for Trapdoor Function and Tdval is the Trapdoor value.

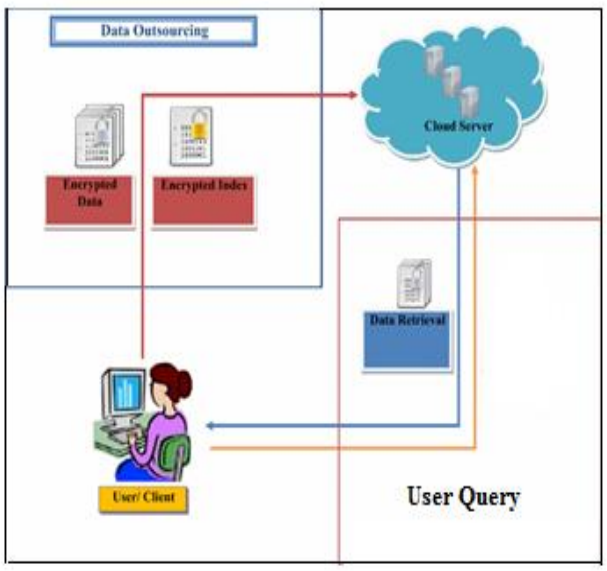

Fig.3: Encrypting Data and Index

Procedure
- Partition insensitive data into blocks.
- Distribute blocks among different cloud storage providers.
- Randomly encrypt some data blocks.
- Maintain storage index information file.
- Encrypt the storage path information using trapdoor function.
- Share trapdoor value to authorized users to decrypt storage index file and to retrieve data.

The storage path information of big data is very small in size and so it is very easy to encrypt, store, decrypt, and share it among authorized users. Also, some random partitions are encrypted making insensitive data blocks more secure. Encrypting random partitions involves: Key generation, Encryption, and Decryption.

Procedure

1. Key Generation

1: Receiver generates Public key and Private Key.
2: Receiver sends public key to all Senders.

2. Encryption

D←Data
Sk←Symmetric Key
Ct ←Ciphertext
ESk←Encrypted Symmetric Key
M←Concatenated Message
L←Length
1: Sender generates Sk.
2: Create ciphertext Ct with ESk.
3: Encrypt and calculate L.
4: M=concatenate(L, ESk,Ct) and return M.

3. Decryption

1: Extract L.
2: Extract ESk.
3: Decode and Decrypt ESk with private key.
4: Decode and Decrypt Ct with Sk.
5: Return D.

4.3. Protecting Public Data

Public data can be accessed freely without imposing security constraints. Public data is also partitioned into blocks and moved to public cloud. Anyway, in some cases public data can be shared only among members of an organization. In such case the previously described storage index encryption methodology can be used to protect public data. The difference is that for protecting insensitive data the storage path information is encrypted along with encrypting some random data blocks. But, for protecting public data only the storage path index is encrypted without encrypting data.

V. EVALUATION AND EXPERIMENTAL RESULTS

Theoretical Analysis

In this work, the computation overhead involved in encrypting and decrypting data is reduced by encrypting only the storage path information without encrypting actual data. The size of storage path information is very small when compared with the actual data and so it is easy to encrypt storage index and to share it among authorized users. In the proposed scheme, the big dataset is partitioned and partition identifier is assigned to each partition in order to map the contents. The partitions are moved to number of storage providers depending on the size and complexity of big dataset. As the data is distributed across number of storage providers, the probability of finding the storage path information of a partition is $p(0<p<1)$. By considering n partitions of data distributed among m storage providers, the probability of finding all the storage path information by an intruder is very less. If the number of storage providers, m is large, the overall probability of acquiring storage path information is p^m, which is very less and negligible.

Security Analysis

The proposed method handles integrity threats and privacy threats that may result in leakage of data to unauthorized users. The usage of proxy re-encryption supports unidirectional delegation, i.e., delegation from A→B does not allow re-encryption from B→A. Also, re-encryption does not need third party interaction and so the process is secure. This work is evaluated with respect to storage overhead and computation overhead. Storage overhead is reduced by the use of hybrid cloud. Usage of private cloud is limited with big data processing as it does not scale well with the growth in big data size. As this approach used private cloud for storing sensitive data and public cloud for insensitive and public data, the storage overhead is less. Public cloud scale well and suits well for large amount of data.

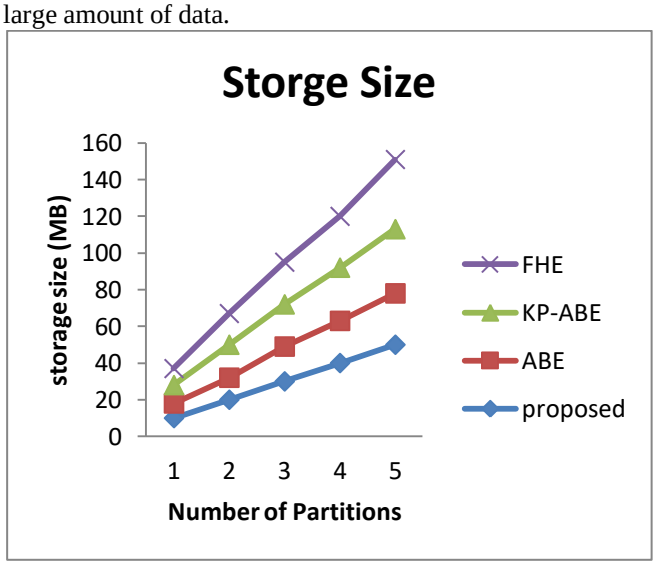

Fig. 4: Storage Overhead

The proposed work is compared with other approaches such as Attribute Based Encryption (ABE), Key-Policy Attribute Based Encryption (KP-ABE), and Fully Homomorphic Encryption (FHE). As these techniques protect data as a whole than encrypting only sensitive data the storage overhead is high. The storage overhead and computation complexity is reduced with the proposed approach.

VI. CONCLUSION

In this paper, a mechanism is provided for using hybrid cloud platform by partitioning big dataset into blocks based on the importance of data. Sensitive data is well protected by implementing proxy re-encryption and by storing data in private cloud. Insensitive and public data can be moved to public cloud by randomly encrypting data blocks and by encrypting storage path information. The analysis results demonstrate that the proposed mechanism is suitable for storing cloud tenants data securely with reduced computation cost and also is effective in data sharing and analytics by the use of hybrid cloud.

REFERENCES

1. Guiyi Wei, Jun Shao, Yang Xiang, Pingping Zhu, Rongxing Lu, "Obtain Confidentiality or/and authenticity in Big Data by ID-based Generalized Signcryption," Elsevier Journal on Information Sciences, 2014.
2. Zhiyuan Tan, Upasana T. Nagar, Xiangjian He, Priyadarsi Nanda, Ren Ping Liu, Song Wang, and Jiankun Hu, "Enhancing Big Data Security with Collaborative Intrusion Detection," IEEE Cloud Computing, 2014.
3. Xiaochun Yun et al., "FastRAQ: A Fast Approach to Range Aggregate Queries in Big Data Environments," IEEE Transactions on Cloud Computing", Vol.3, No.2, April/June 2015.
4. Kaitai Liang, Willy Susilo, and Joseph K. Liu, "Privacy-Preserving Ciphertext Multi-Sharing Control for Big Data Storage," IEEE Transactions on Information Forensics and Security, vol.10, no.8, 2015.
5. Melbin J Reena, A. ShajinNargunam, "A Review on Cryptographic Approaches for Secured Processing of Big Data" IJCTA, 10(03), pp. 73-79, 2017.
6. C. Moretti, K. Steinhaeuser, D. Thain, N.V. Chawla, " Scaling up classifiers to cloud computers," Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pages 472-481, Washington, DC, USA, 2008.
7. JinboXiong, Ximeng Liu, Zhiqiang Yao, Jianfeng Ma, Qi Li, KuiGeng, and Patrick S. Chen, "A Secure Data Self-Destructing Scheme in Cloud Computing", IEEE transactions on cloud computing, vol.2, no.4, 2014.
8. J.Han, M.Kamber, and J.Pei. "Data Mining: Concepts and Techniques," San Mateo, CA, USA: Morgan Kaufmann, 2006.
9. Carlos E. Otero, Adrian Peter, "Research Directions for Engineering Big Data Analytics Software", IEEE, 2015.
10. Boyang Wang, Baochun Li, and Hui Li, "Oruta: Privacy-preserving Public Auditing for Shared Data in the Cloud," IEEE Transactions on Cloud Computing, vol.2, no.1, 2014.
11. Chang Liu, Jinjun Chen, Laurence T.Yang, Xuyun Zhang, Chi Yang, Rajiv Ranjan, and RamamohanaraoKotagiri, "Authorized Public Auditing of Dynamic Big Data Storage on Cloud with Efficient Verifiable Fine-Grained Updates," IEEE transactions on parallel and distributed systems, 2014.
12. Muhammad MazharUllahRathore et al. "Real-Time Big Data Analytical Architecture for Remote Sensing Application" IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol.8, No.10, October 2015.
13. JoonsangBaek, QuangHieu Vu, Joseph K. Liu, Xinyi Huang, and Yang Xiang, "A Secure Cloud Computing based Framework for Big Data Information Management of Smart Grid," IEEE transactions on Cloud Computing, vol.3, no.2, 2015.
14. Rongxing Lu, Hui Zhu, Ximeng Liu, Joseph K. Liu, and Jun Shao, "Toward Efficient and Privacy-Preserving Computing in Big Data Era", IEEE Network,2014.
15. MatturdiBardi, Zhou Xianwei, Li Shuai, and Lin Fuhong, "Big Data Security and Privacy: A Review," China Communications, supplement no.2, 2014.
16. Amy Xuyang Tan, Valerie Li Liu, Murat Kantarcioglu, BhavaniThuraisingham, "A Comparison of Approaches for Large-Scle Data Mining," Technical Report, The University of Texas at Dallas, August 2010.
17. Han Hu, Yonggang Wen, Tat-Seng Chua, and Xuelong Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," IEEE Access, 2014.
18. Xinhua Dong, Ruixuan Li, Heng He, Wanwan Zhou, ZhengyuanXue, and Hao Wu, "Secure Sensitive Data Sharing on a Big Data Platform," Tsinghua Science and Technology, ISSN 1007-0214 08/11 ,pp 72-80, volume 20, number 1, 2015.
19. Cheng Hongbing, RongChunming, Hwang Kai, Wang Weihong, and Li Yanyan, "Secure Big Data Storage and Sharing Scheme for Cloud Tenants," China Communications, 2015.
20. Mohammad UbaidullahBokhari, QahtanMakkiShallal, "Evaluation of Hybrid Encryption Technique to Secure Data during transmission in Cloud Computing", International Journal of Computer Applications, Volume 166, No.4, May 2017.
21. AbidMehmood, IynkaranNatgunanathan, Yong Xiang, GuangHua, Song Guo, "Protection of Big Data Privacy", IEEE Access, Volume 4, 2016.
22. G. Wang, F. Yue, and Q. Liu, "A secure self-destructing scheme for electronic data", Journal of Computer and System Sciences, vol. 79, no. 2, pp. 279–290, 2013.
23. L. Zeng, Z. Shi, S. Xu, and D. Feng, "Safevanish:An improved data self-destruction for protecting data privacy", Proc. 2nd Cloud Computing International Conf., Indianapolis, USA, 2010, pp. 521–528.
24. L. Dong, Y. Zhuang, Y. Gao, and Y. Bu, "Research on realtime trigger system for sensitive data safe destruction", Journal of Chinese Computer System, vol. 31, no. 7, pp. 1323–1327, 2010.
25. J. Qin, Q. Deng, and J. Zhang, "Design of multi-grade safety data destruction mechanism of HDFS", Computer Technology and Development, vol. 23, no. 3, pp. 129–133, 2013.
26. F. Zhang, J. Chen, H. Chen, and B. Zang, "Lifetime privacy and self-destruction of data in the cloud", Journal of Computer Research and Development, vol. 48, no. 7, pp. 1155–1167, 2011.