ATLANTIS PRESS
Advances in Computer Science Research, volume 70
2nd International Conference on Mechatronics Engineering and Information Technology (ICMEIT 2017)

Data Security and Privacy Preservation in Big Data Age

Mingsheng Yin
International School Beijing University of Posts and Telecommunications, Beijing, 100878, China
mingshengyin@foxmail.com

Keywords: data security, privacy preservation, big data; cloud computing, data encryption.

Abstract. The advent of the big data age has provided great convenience for people in some ways, while it appeared some drawbacks recently, such as a large number of users' data being leaked and people's privacy being exposed to the public. It brought great trouble to those whose information has been exposed. So data security and the preservation of people's privacy is an important issue that is urgently needed to solve in big data age. This paper gives a brief exposition on the aspects of the concept of big data, the system architecture of data protection, the research content and technical methods of privacy preservation. In order to review the research progress of current data security and privacy preservation and try to find out the shortcomings and the future research direction of information security technology, which could make people enjoy better the convenience of the information age.

1. Introduction

With the rapid development of Internet technology, the volume of global data shows explosive growth. Data mining technology aggregates these data previously unassembled and discover more quickly and accurately valuable information from massive, incomplete, noisy, fuzzy, random, large databases [1]. By analyzing the information and then make inductive reasoning, from which to explore the potential model to help people make the right decision. However, science and technology is a double-edged sword, in the human life for the great convenience, the hidden dangers behind big data cannot be underestimated. With the virtualization, cloud computing and other new technologies widely being used, Internet privacy leaks are always coming about [2]. How to be able to enjoy the convenient life in big data age and effectively avoid the threat posed by it, has become the focus of the current study.

Whether it is to read the website or shopping site, there are numerous contents of the recommendation based on the stay time and the contents that the user had been browsing the pages, after the analysis of that the user may be interested in, which largely facilitates the user on searching and selection. But behind the access to convenient and personalized service, it to some extent exposed the user's privacy. When the user using the Internet, the information is unknowingly recorded. When they using mobile phones, the objects and talk time, even the location of calling are recorded. When you send a message or share a photo, the Internet operator can get the user's preferences. With the rapid development of data acquisition technology, personal interests, physical characteristics and other privacy information can be easily obtained without notice of the user. The large number of detailed data generated from big data age can be used to describe the behaviors of various objects, societies and even the whole environment. By analyzing these data, we can greatly reduce the complexity of society, improve people's understanding of the world, and enhance the ability of modifying the world and help people make important decisions. If this information is effectively used, it will bring a lot of convenience to human life, but if it's unlimited or even malicious use, the bad consequences will be incalculable [3].

2. The concept of big data

Big Data refers to a large, complex set of data that cannot be extracted, stored, searched, shared, analyzed, and processed with existing software tools. The generation of big data is the product of the

development of the Internet, the type of data is also varied, including text, video, audio, pictures, geographic information and network log, etc. The data comes from a variety of information generated by people doing Internet activities, various types of documents, databases and logs produced from computer system and mobile phone system, a variety of information channels collected from all kinds of data devices. Big data is the forefront of digital analysis technology, whether in life or in the development of science and technology, big data technology have been widely used [4].

3. The system architecture of data privacy preservation

In big-data cloud computing, mobile cloud services can provide flexible storage, computing and other resources on demand anytime, anywhere, combined with the current research and application of mobile cloud services, its architecture can be divided into IaaS, virtual layer, PaaS and SaaS and mobile terminal, which is shown in Figure 1.

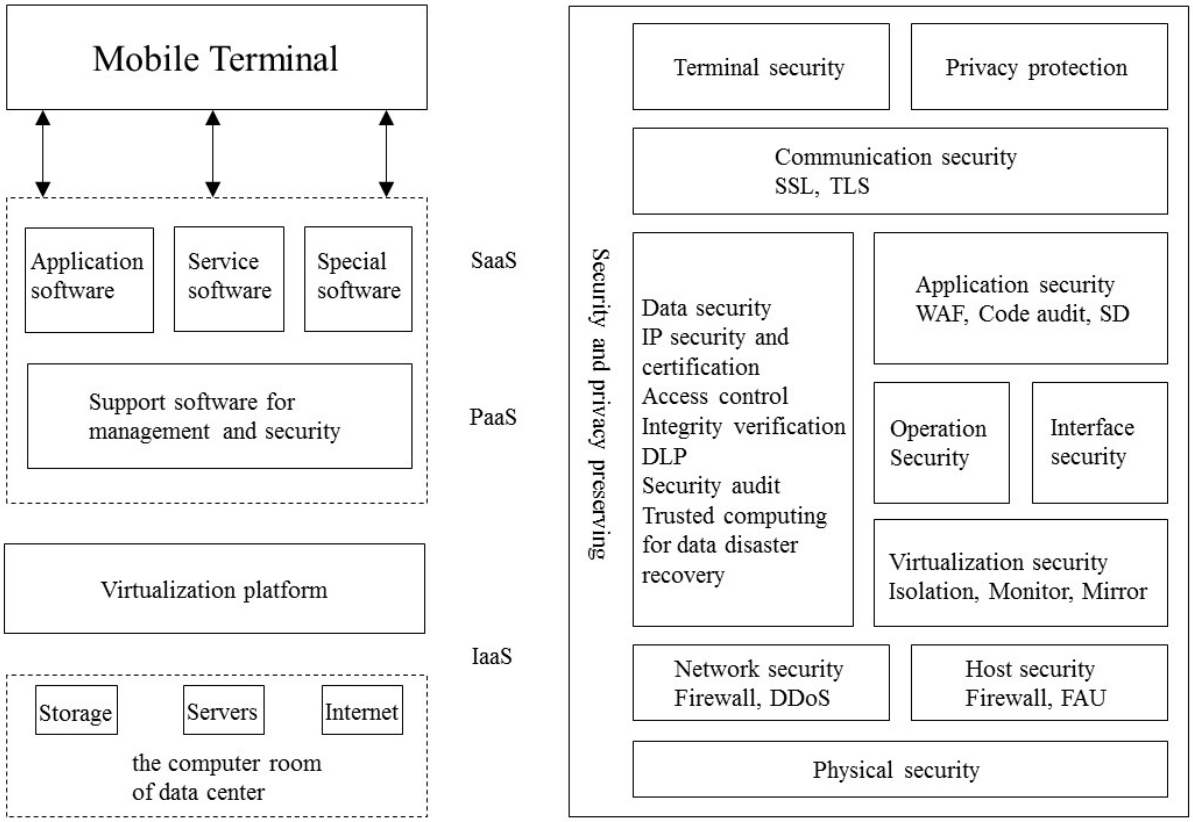

Fig. 1 System architecture of security and privacy preservation

The security threats faced by IaaS are mainly physical security, network security and host security. Virtual layer mainly has virtualization security issues, which generally solved by the mirror reinforcement, configuration management, virtual machine attack protection and other measures. While there are application security, operational security, interface security and data security existing in PaaS and SaaS. The mobile terminal includes terminal security and privacy preservation.

4. The main research fields of data privacy preservation

The issues of privacy preservation are accompanied with data applications. In the field of statistics, privacy-preservation issues are the first concern [4]. At present, the main research fields of privacy preservation is shown in Table 1.

The problem of privacy preservation is determined by the different privacy requirements under practical applications. Universal privacy preservation technology is committed to protecting data privacy at a lower application level, which is generally achieved through the introduction of statistical models and probability models. While the privacy preservation technology for data mining mainly deals with high-level data applications, which is to exploit the characteristics of different data-mining operations. The principle of data release based on privacy preservation is to provide a privacy

protection method that can be used in all types of applications, so that the design of the privacy preservation algorithm is also versatile on this basis.

Table 1. Research Fields of privacy preservation

Research Fields	Example
Universal privacy preservation technology	Perturbation [4] Randomization [5] Swapping [6] Encryption [7]
Privacy preservation technology for Data Mining	Association Rule Mining [8] Classification [9] Clustering [10]
The principle of data release based on privacy preservation	k-anonymity [11] l-diversity, m-Invariance, t-Closeness [12]
Privacy preservation algorithm	Anonymized Publication [13] Anonymization with High Utility [14]

5. The classification of data privacy preservation technology

No privacy preservation could apply to all applications, so the privacy-preservation technology is classified into 3 groups:

(1) Data-distorting technology, it is a method of distorting sensitive data while keeping some data or data properties unchanged. For example, adding noise and swapping is used to process the raw data. But it is required to ensure that the processed data should still maintain certain statistical properties in order to perform the data mining operation.

(2) Data-encryption technology, it is the use of encryption technology in the data mining process to hide sensitive data, which is mainly applied to distributed application environment, such as secure multiparty computation (SMC) [15,16].

(3) Limited-release technology, it is conditional release of data based on the specific circumstances. Such as not publishing some field values of the data, data generalization [11].

In addition, for many other methods [2-3], because of the integration of a variety of technologies, it is difficult to classify then into some of the above. But during making use of some of the advantages of one technology, it will inevitably bring some defects. Data-distorting technology has a relatively high efficiency, but there is a certain degree of information loss. Data-encryption technology is just the opposite, it can guarantee the accuracy and security of the final data, but the calculation of overhead is relatively large. While the advantage of limited-release technology is to ensure that the published data must be real, the number of published information will be a certain loss.

Each type of privacy preservation technology has different characteristics. In different application requirements, their scope of application, performance, etc. are not the same. Comparative result is shown in Table 2.

Table 2. Comparative result of different privacy preservation technologies

Representative technology	Universality	Computational cost	Data loss	Privacy preservation
Data-distorting technology	Medium	Low	High	Medium
Data-encryption technology	High	High	Low	High
Limited-release technology	High	Medium	Medium	High

In the mobile cloud environment, it is needed to consider using different privacy preservation technologies according to different needs. At present, research on privacy preservation in mobile cloud services focuses on the establishment of privacy policy from the perspective of cloud service providers, which ignores the dynamic nature of privacy-preservation requirements. The security mechanisms such as encryption and access control strategies could only protect the user direct privacy. However, there are many factors that may lead to privacy disclosure, and the privacy-preservation needs of terminal-user vary widely. Relying on traditional security verification and management strategies can not meet the user's potential privacy-preservation needs.

6. Conclusion

Big data makes human life becomes convenient and efficient, but the frequent issue of privacy disclosure alarms people who are enjoying this convenience. Privacy and security issues have been paid close extensive attention. In this paper, the research progress of data security and privacy preservation under big data age is introduced from the aspects of the architecture of data privacy preservation, the research fields of data privacy preservation, and the classification and comparison of privacy preservation technology. Although data security and privacy preservation research has made some progress, there are a lot challenges in the future. The author believes that future research work can focus on the following aspects:

(1) The contradiction between big data cloud computing service quality and privacy preservation. When users use cloud-based services, they need to provide some of their own data, the higher the accuracy of the data, the better the quality of service, but the lower the degree of privacy. The balance between privacy and quality of service is a challenging issue.

(2) The security and privacy preservation of dynamic data. Mobile devices in the social network produce a lot of dynamic data in real time, data patterns and data content are changing at all times, and the existing privacy preservation technology is mainly based on static data sets. In view of the attackers making use of accumulation and relevance of the data to extract, integrate, analyze and mine user privacy data, how to achieve security and privacy preservation of dynamic data in the cloud computing environment will be more challenging.

(3) Establish a sound security technology framework, service safety standards and its evaluation system. The establishment of safety guidance standards and its evaluation technology system is also an important pillar to realize the security of cloud computing services.

With the continuous development of privacy preservation technology, I believe people could freely enjoy the life of intelligent data era.

References

[1] Agrawal R, Srikant R. Privacy-preserving data mining. Research Journal of Applied Sciences Engineering and Technology, 2003, 9: 616-621
[2] Miklau G, Suciu D. A formal analysis of information disclosure in data exchange. Journal of Computer and System Sciences, 2007, 73: 507-534
[3] Machanavajjhala A, Gehrke J. On the efficiency of checking perfect privacy. Acm Sigact-sigmod-sigart Symposium on Principles of Database Systems, 2006: 163-172
[4] Adam N, Wortmann J C. Security-control methods for statistical databases: A comparison study. ACM Computing Surveys, 1989, 21: 515-556
[5] Warner S L. Randomized response: A survey technique for eliminating evasive answer bias. The American Statistical Association, 1965, 60: 63-69
[6] Fienberg S E, Mcintyre J. Data swapping: Variations on a theme by Dalenius and Reiss. Privacy in Statistical Databases: Casc Project International Workshop, 2004, 52: 14-29
[7] Pinkas B. Cryptographic techniques for privacy-preserving data mining. Acm Sigkdd Explorations Newsletter, 2002, 4:12-19
[8] Evfimievski A, Srikant R, Agarwal A, Gehrke J. Privacy preserving mining of association rules. Information Systems, 2004, 29:343-364
[9] Wang K, Yu P S, Chakraborty S. Bottom-up generalization: A data mining solution to privacy protection. IEEE International Conference on Data Mining, 2004:249-256
[10] Vaidya J, Clifton C. Privacy-preserving k-means clustering over vertically partitioned data. Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2003: 206-215
[11] Sweeney L. k-anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10: 557-570
[12] Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. IEEE International Conference on Data Engineering, 2007: 106-115

[13] Aggarwal C C. On k-anonymity and the curse of dimensionality. International Conference on Very Large Data Bases, 2005: 901-909
[14] Kifer D, Gehrke J. Injecting utility into anonymized datasets. Acm Sigmod International Conference on Management of Data, 2006: 217-228
[15] Yao C C. How to generate and exchange secrets. Symposium on Foundations of Computer Science, 1986, 10: 162-167
[16] Clifton C, Kantarcioglu M, Vaidya J, Lin X, Zhu M Y. Tools for privacy preserving distributed data mining. Acm Sigkdd Explorations Newsletter, 2002, 4: 28-34