Privacy Protection under Big Data Technology Environment

LI Weichao[1,2]
[1]Zhengzhou University of Aeronautics, Zhengzhou 450046, China
[2]Collaborative Innovation Center for Aviation Economy Development, Zhengzhou 450046
E-mail:liweichao@zua.edu.cn

ZHAO Haixia[1]and ZHANG Liuqi[1]
[1]Zhengzhou University of Aeronautics, Zhengzhou 450046, China

Abstract—In today's society, Big Data exists everywhere. The paper makes analysis on characteristics, sources and framework of Big Data. Big Data brings us with many benefits as well as challenges to our personal privacy, so the author explains the reasons why personal privacy is disclosed and the ways to reveal personal privacy. The author proposes three pieces of strategies of privacy protection in the end. Firstly, perfect legislation and rigorous supervision are regarded as the fundamental of privacy protection mechanism. We are in urgent need of a comprehensive and unified Internet privacy protection legal system. Secondly, advanced technology is an important aspect of strengthen the privacy protection, and this paper have introduce some key technologies to privacy protection. Last but not least, it needs cooperation between users and enterprises in order to construct a protection system. On the one hand, users should raise awareness of privacy protection. On the other hand, enterprises should give users more control space for personal data.

Index Terms—big data; privacy protection technologies; information security

I. INTRODUCTION

Nowadays, the rapid development of information technology has led to a surge in the amount of Internet data. According to incomplete statistics, the amount of the information shared by users per day is more than 2 billion, the number of Twitters will be more than 1.3 million per hour, the number of users to search Baidu has reached 1.5 million. At the same time, it also produces a lot of data in the financial industry, service industry, government information, scientific computing and other fields all the time. In 2012 the amount of global information has reached 2.7ZB, and in 2015, it is expected to reach 8ZB [1].

Big Data era has come, and it fills our daily life everywhere. Big Data contain a lot of useful data, and enterprises can get a lot of economic value and benefits from mining these useful data. Big Data technology is making progress rapidly, and some institutions or enterprises can dig, collect or collate personal privacy without users' permission, which is a great invasion of personal privacy [2]. Now in this data-flying era, everyone seems to be naked shown in front of the public. This causes a great concern on privacy issues in our community, and users begin to be anxious about their privacy. Some foreign countries have made research on privacy protection. And yet, how should we protect personal privacy in China? On the basis of analyzing the present research situation of Big Data, this paper describes the security challenges of Big Data, and put forward some methods to construct a system of personal privacy protection.

II. OVERVIEW OF BIG DATA

A. Sources and characteristics of Big Data

Big Data is aggregate of large scale of online data. It has characteristics of huge volume, high velocity, variety and value, which is called "4V" of Big Data [3]. The source of Big Data is various, such as studying, working, daily life, and entertainment. The Big Data are generated all the time. Besides, through the chat tools, the online shopping, some of our private data are disclosed in forms of texts, images and video. Therefore, some enterprises dig these valuable data to get huge commercial interest, which is an invasion to personal privacy to some extend.

B. Technical framework for Big Data

Big data processing involves data acquisition, preprocessing, analysis and interpretation, etc. Figure 1 is a schematic diagram of the related technology.
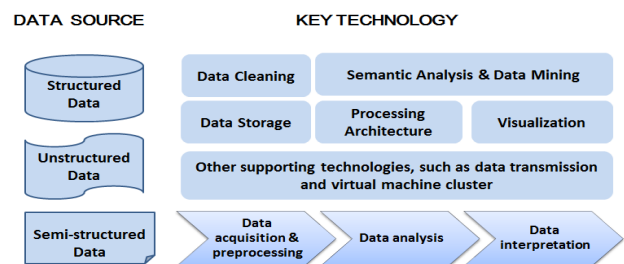

Figure 1 Technical Framework for Big Data

1) Data acquisition and preprocessing. With the rapid development of science and technology, computer technology and Big Data processing technology have also been greatly developed and data are accessed to in a variety of ways, business processing in people's daily life and scientific research have accumulated a lot of data. These data contain a wide variety of data from the server background, text, pictures, video, and other activities made by people. Because of the different information structure characteristics, data can be divided into structured data, unstructured data and semi-structured

data. To data with different structural characteristics, there is a huge gap in the process of acquisition and preparation. Different mode of acquisition should be taken to different structural characteristics. First, the data in the user information table, the list of goods and others, which can be expressed by figure and have relational features, are structured data. To structured data we generally adopt the relational database information acquisition model. The data that cannot be expressed along with figure or structure is unstructured data, such as text information, pictures, videos and so on. To them we generally adopt file system resource acquisition module. The semi structured data are the data set containing both characteristics. We generally adopt the network information acquisition mode and other information resource acquisition modes. After data acquisition, the data should be preprocessed, which is also a process of useful information acquisition. In this process, a large amount of acquired data are integrated and processed, and then are stored in DSM data warehouse in a certain way.

2) Data analysis. After data acquisition and preprocessing, we should analyze the data, which is a very important step, and also the key to Big Data technology. Through data analysis we can discover a large number of useful information implied in the data. According to different level, Data analysis is divided into 3 categories: query & index, semantic analysis & data mining, processing framework. Index is a connection relationship in the database, it can improve the speed of our query data, just like the role of the directory in a book. Simply saying, data mining is the process of sorting out and analyzing data from Big Database. But there is no uniform data mining method because of the huge types of data, not unified semantic analysis. Hadoop multidimensional analysis platform is a present popular architecture, which is divided into four modules. Each module has its own role, and the core module is a parser to transfer the multi-dimensional analysis language to the MapReduce, and can scan mass data.

3) Data interpretation. Visualization and man-machine interaction is the basic way of data interpretation. It is easy to understand the data interpretation is an interpretation of the results of the data analysis so that we can better understand the meaning of data. With the continuous optimization of electronic display wall, multimedia equipment and other hardware facilities, we can achieve visualization of Big Data through the dimension degradation. Compared with foreign countries, the visualization of massive data needs further development in our country. However, there are great limitations to visualization because of our personal reasons. Therefore, man-machine interaction technique will become the core technology of data interpretation where we should have a further study.

4) Data transmission and virtual cluster. Big Data application is mainly on acquisition, integration, analysis, interpretation of the data, and is data-focused. But the data transmission process is also a very important step that affects the whole follow-up data work. In addition, the virtual cluster has many advantages such as structural diversity, easy integration, and small cost. It will also become the mainstream of the popular cluster approach

in dealing with Big Data, but its function and performance need to be further developed.

III. CHALLENGES TO THE SAFETY OF PERSONAL PRIVACY BROUGHT BY BIG DATA

The Big Data age has two facets. On the one hand, it brings a lot of convenience to our life; On the other hand, it poses a threat to our personal privacy. The huge economic profits and market value brought by Big Data have attracted many enterprises [4]. The Prism Event broken out earlier has widely attracted the attention of people from all over the world. Different from the previous information age, the emergence of Big Data has brought huge challenges to the safety of personal privacy.

A. Reasons why personal privacy is disclosed in the Big Data age

The internet age has arrived where electronic technology is developing with an ever-increasing speed, the hardware system and the ubiquitous user smart computers generate enormous data every day. Our data are increasing with an exploding rate. Every minute, every second, there are many data being produced. Through deep digging, enterprises could easily get these data which contain vast personal information of computer users, whose individual characteristics could be gained through data analysis and data searching. These data have great commercial value for enterprises, helping enterprises formulate strategies to achieve higher profits. Nevertheless, getting the data in such a way greatly violates the personal privacy of users. For example, there are plenty of users searching with key words through the searching engine of Baidu. Baidu Inc., through the integration, processing and digging of these data, can find out the data which catch the most attention from users, and then promote service to users, which brings them a rich reward every year. However, the recycle that Baidu Inc. makes on the vast data created by users has obviously led to the violation to personal privacy. Personal data are a part of Big Data. Enterprises dig into the Big Data, use system software and date processing technology, to integrate useful user information. On the one hand, this is information utilization for enterprises and can bring them huge economic benefits; However, on the other hand, we do not agree to this, because users are not familiar with or even do not know about the situation when their information is dug and achieved. Our daily information, personal likes and dislikes and contact ways are exposed on the internet to enterprise technicians to make better decisions, which, to a large degree, has violated the users' privacy. With small cost but good awards, this means of information utilization is widely adopted by many companies and enterprises who gain commercial value through digging into users' data information. The protection of personal privacy has becoming a worrying concern.

B. Ways to reveal personal privacy

Some people say that in the information age, people are deprived of any privacy just like the goldfish living in the bathtub. In the contemporary society, all kinds of privacy and breaking news constantly pop out from the

internet, including stars' phone number, their life situation, and even the check-in records of many celebrities. The protection of personal privacy is greatly concerned. Then what are the ways to reveal personal privacy? The following is a simple introduction [5].

1) Daily registers. In our daily life, we are requested to fill out our true information when we need accommodation, traveling, shopping and hospital registration. The recorded information has some certain functions after being integrated and processed and could probably be exchanged as a commodity.

2) Public exams and job seeking.When we are required to fill out personal information for different kinds of employment opportunities on the internet with our phones, computers or other softwares and submit them to the employment websites, our personal privacy has already been exposed without attention, and these information are most likely to be either actively or passively stolen by some attacking softwares.

3) Third party attacking software.This is a kind of virus software. The virus may invade the user's individual computer or some government website systems, searching, copying and stealing their data information, which greatly violates their personal privacy. They carry out devious conducts by illegally using these information. Some softwares and hackers even steal users' financial information such as bank account and passwords, which causes huge economic losses for users.

4) Users excessively publishing personal privacy. Some internet users do not pay attention to their own information security. When surfing the internet, they will inadvertently disclose their personal information to others. Their privacy is thus difficult to get protected. For example, when chatting on WeChat, chatters will make their own personal data public, which makes it easier for lawbreakers to use.

5) Phishing websites. Phishing websites are disguised as shell websites on the internet such as electronic banks, online shops, to steal personal bank card information, passwords and other users' personal information. These lawbreakers tamper with the website URL and make their websites look like normal websites, then promote their websites in many ways. Sometimes, they will use some gifts to attract netizens to fall for them. Internet users are massive. Some will fall for these websites without enough awareness and get their personal privacy stolen by the phishing websites, and even suffer from tremendous economic losses

C The issue of privacy protection already arousing users' concern

One of the characteristics of Big Data is the large amount of data information, which can expound the nature of some issues in a broader sense, thus giving enterprises a more sensible and accurate direction. For example, websites such as Baidu, 360buy and Facebook gain huge economic profits by analyzing and digging into customers' purchasing patterns, personal preferences, friends circle, their caring aspects and other information. In the Big Data age, every single information is being records when users visit their friends' updates, post

twitter messages and make online shopping, which virtually brings great security challenges to users' privacy. At present, people's consciousness for privacy protection is gradually awakening. According to some certain investigation made among more than 10 thousand people from many countries all over the world, most people being investigated think their online records and personal privacy unsafe and have the tendency to be disclosed. Only 14% of the people think that their online data will be reasonably and publicly used by enterprises and companies. 69% of the respondents suggest that they will not worry about their privacy being leaked out while they surf the internet if the users' online information will not be recorded in the internet age. This investigation indicates that most users have already had the protection awareness of their privacy.

In recent years, incidents of people's personal privacy being exposed on the internet have emerged continuously one after another, and the hazards are also very obvious and great. Users' privacy safety faces unprecedented threat. But the relevant technology to protect people's privacy in the Big Data age urgently needs improvement so as to adapt to the development of times.

IV. PRIVACY PROTECTION STRATEGIES UNDER BIG DATA ENVIRONMENT

A. Perfect legislation & strengthen supervision

Certainly, it is important to understand the Big Data flow when we in the use of Big Data, meanwhile we must realize the significance of legal system. In the era of big data, it will lead the Internet service provider to misuse Big Data and violate citizens' basic rights if we admit that we could protect our privacy effectively by the existing legal norms.

1) Data privacy, has become a new concept of privacy protection in Big Data era. In terms of our country, the legislation of information is still not perfect, although China's criminal law has to legalize "violate citizens personal information crimes", the relevant administrative regulations and department rules is still more extensive. In order to deal with the impact of Big Data era, privacy protection in the 21st century should be transferred from people-oriented to data-oriented as the train of thought. Nowadays, the information conveyed by data undoubtedly constitutes the important content of modern society citizens' privacy, so countries under the rule of law ought to provide full security for citizens' data privacy. In this respect, the United States has provided experience for us. The United States congress passed the Privacy Act as early as 1974 establishing the basic principles of data privacy. As for how to manage the data related to personal identity privacy, the United States congress has passed the Electronic Communications Privacy Act, the Computer Check and Privacy Protection Act, the Federal Information Security Management Act, etc.

2) We are in urgent need of a unified Internet privacy protection legal system. In addition to the relevant laws

on the traditional privacy protection, the laws in our country about the Internet privacy involved the Decision on Maintaining Internet Security published by NPCSC, the Telecommunications Regulations published by the State Council, the Telecom and Internet Users' Personal Information Protection Sets published by MIIT, and the Measures for the Administration of Online Transactions published by SAIC of China, etc. However, some are not so feasible, and don't have high legal status or effectiveness. Overall, the relevant laws, regulations and rules relating to privacy protection lack of cohesion, and sometimes they are paradoxical. Big data is the trend of the development of science and technology, and privacy leakage is also a difficult problem in front of human beings. Our country ought to not only enact keeping pace with the times, but also insure the laws more linked and systematic. In order to fundamentally ensure the safety of personal privacy, we are in urgent need of a unified Internet privacy protection law system [6].

3) More regulation, serious accountability should keep pace with perfecting legislation.Our country existing law system clearly requirements that enterprises should collect and use users' data before asking for owners' permission. Government departments must strengthen supervision and management while ensure enterprises to perform its obligations. On the one hand, government should set up specialized agencies as soon as possible, so that all the obligations of data collection and using will be detailed. On the other hand, government should establish an efficient investigation and intervention mechanism, so that government could intervene for investigation more quickly while dealing with consumer complaints.

B. Technology research & technology application

Technology is an important aspect of strengthen the privacy protection. A report from WEF (World Economic Forum), published in 2014, proposed to rely on technology to protect privacy, and technology should be an important method of privacy protection. It is recommended to strengthen technology research on privacy protection while supporting the development of Big Data collecting or processing and other technology, in order to ensure privacy from the technical level. In this chapter, we will explain some key technologies.

1) Data publishing anonymous protection technology. Anonymous protection is the most common and one of the most important technologies in the privacy protection (Anonymization is one of the primary techniques realizing privacy protection in data dissemination environment.), which is often used to deal with relational data. Under the environment of Big Data, the attacker can not obtain information from an anonymous data, but when the data are huge, the attacker will have chance to crack the kind of anonymous method and data will be lost. So the technology needs to be improved and refined constantly.

2) Data watermarking technology. Data watermarking technology is a kind of technique to hide information, and it labels specific identity on original data (text, pictures, and videos) to ensure that the information is true and the original data is not tampered. These digital watermarks are often interspersed in the data information and not easy to be found, and will not affect the information originally expressed by data.

3) Anonymous protection technology of social network. In today's Internet era, social chat has become an essential part of daily life. There is great need to protect privacy in Social network. According to the different hidden content, it is generally divided into nod anonymity and edge anonymity. Social network anonymous technology is faced with serious problems, and hackers can analyze not anonymous data information to infer the users' anonymous information. This conjecture depends on the connection between the anonymous information and other information. The less the relationship is, the more difficult it is to infer the anonymous information. The future technology will focus on anti speculation, and anonymous social network protection technology will make progress quickly.

4) Data tracing technique. At the beginning when the Big Data was put forward, the data tracing technology has been explored accompanied with the development of the database, the fundamental purpose is to assist staff to better position the "origin" of each datum in data warehouse. For example, we make it clear that the information is calculated and obtained based on which part and which form the data come from. And depending on these information we can easily test the accuracy of the final conclusion, or update the new data with a very little effort. As we all know, the most essential computing method of data traceability is the marking method. After a long-term development, the definition of deep level is divided into where- and why- two classification, finally according to the characteristics of their own category one focuses on computing method, the other focus on the source of data. In addition to various data types in the database, it also includes the XML analysis of data, data flow and data tracing technology to the undetermined data. Besides, this special technology can also be used to trace and recover the important lost information or the destroyed document. We reasonably believe that the digital tracing technology can produce a great usage in information security in future.

5) Role mining. Nowadays, network access control is more popular, and the role based access control (RBAC) is one of the most popular access control techniques. This technology is also a kind of important protection technology in information security. RBAC is rather outstanding in dealing with a large number of access control problems. By introducing the concept of role very early, it improves the security performance of the system and simplifies management by combining role and permissions [7].

C. Raising awareness & guaranteeing right.

1) From users' perspective, they should raise awareness of privacy protection. In Big Data era, users are not only the consumers of data, but also the producers. Users are obliged to submit their own data in order to enjoy personalized and precise services. But

users should know what the rights they have about personal data, and be cautious of excessive data acquisition and data abuse. At the same time, in the process of using data service, users need to preserve evidence in important link, and maintain their own rights and interests by legal means if necessary.

2) From enterprises' perspective, they should give users more control space for personal data. Gathering relevant data is inevitable when enterprises are providing users with accurate and personalized services. However, the excessive collection of data and data abuse will cause the user's disgust. Therefore, the enterprises should give users more control of their personal data and more options, protect the users' right to know, and use data resource reasonably within the scope of the law. Enterprises, for example, allow users to choose whether to use personal data for advertising, personal data in Internet enterprise shelf life, etc [8].

CONCLUSION

Under the rise of the cloud technology, despite of more new security incidents produced in the Big Data era, it is undeniable that it itself is also the best way to cope with difficulties. In this article, by analyzing the personal privacy challenges caused by Big Data era, the author introduces some key technologies of privacy protection. Compared with foreign countries, there are significant gaps for the technology of Big Data security and privacy protection. But we firmly believe that our technicians are also trying to develop the relevant technologies, and the relevant policies and laws of privacy protection in Big Data era will also be improved. Big Data security and privacy protection will be better resolved in the future.

ACKNOWLEDGMENT

REFERENCES

[1] Z.P. Zhang. "Big Data security and privacy protection", *Information Security and Technology*.vol.1,2015,pp.11-12.
[2] X.L. Yan. "Big data analysis and personal privacy protection",*China Information Security*.vol.3,2014,pp105-107.
[3] L.N. Jiang. "Behind the big data information security applications",*China Information Security*.vol.8,2013, pp.105.
[4] Y.H. Liu, Zhang Y. Tie,X.L. JIN,and X.Q. Cheng. "Personal privacy protection in the era of Big Data,"*Journal of Research and Development*.vol.52,2015, pp229-247.
[5] X.Y. Gu."Network information security issues in Big Data era environment", *China Management Informationization*. vol.18, 2015, pp.163.
[6] X.S. Qiao, Z.D. Zhi, B. Wang,"Big Data,personal privacy-How should we protect our privacy in Big Data era?"*Friend of Science Amateurs*.vol.1,2016,pp.10-15.
[7] D.J. Feng, M. Zhang and H. Li."Big Data Security and Privacy Protection",*Chinese Journal of Computers*.vol.37, 2014, pp.246-258.

LI Weichao, PhD, associate professor, specializes in digital information resource management.
E-mail:liweichao@zua.edu.cn