ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(24), DOI: 10.17485/ijst/2015/v8i24/80150, September 2015* ISSN (Online) : 0974-5645

Data Security Scheme for Multiple Attribution Information in Big Data Environment

Yoon-Su Jeong[1], Yong-Tae Kim[2*] and Gil-Cheol Park[2]

[1]Division of Information and Communication Convergence Engineering, Mokwon University, Republic of Korea; bukmunro@mokwon.ac.kr

[2]Division of Multimedia Engineering, Hannam University, 133 Ojeong-dong, Daedeok-Gu, Daejeon - 306-791, Republic of Korea; ky7762@hnu.kr, gcpark@hnu.ac.kr

Abstract

Recently, as the installation and deletion of not only the Internet but also applications have become easy without any time or place constraints through 3G networks as well as other diverse interfaces such as Wi-Fi and Wibro, the amounts of data are increasing and the kinds of data are being diversified. However, socially large repercussions are arising because methods of protecting user data are insufficient. In the present paper, a security scheme is proposed that will link big data to attribution information to enable users to be provided with big data services from different networks in real time. The proposed scheme inserts attribution information at the beginning and end of contents for provision of big data contents services in real time so that both authentication and non-repudiation can be provided for the contents even if one of the two signatures is lost. In addition, the proposed scheme enhanced the security of contents data by transferring hash chain values to prevent attribution information from being unnecessarily exposed to any 3rd party.

Keywords: Big Data, Data Security, Hash Chain, Multiple Attribute

1. Introduction

Recently, big data services that store the data processed in cloud environments in heterogeneous devices so that the data can be easily used in different network environments are in the limelight[1,2]. In particular, the development of big data technologies characterized by the creation, collection, analysis, and expression of diverse kinds of large scaled data enables us to more accurate predict the diversified modern society to have it operate efficiently, provide, manage, and analyze customized information for each individualized member of modern society, and realize technologies that have been impossible in the past[3–5].

As big data are propagated and activated, socially large repercussions are arising because of those security threats that have been occurring in existing PCs and newly occurring security threats. Big data are characterized by the ability to use diverse programs and data and continuously add and delete programs because high-capacity memories are adopted and an operating system is loaded. In particular, big data users have been easily exposed to important malware attacks since the related virus appeared in 2004 because the centralized management of big data can be easily exposed to hackers to become a target[6,8].

Big data present the possibility to provide valuable information to society and mankind across all areas including politics, society, culture, and scientific technologies and their importance is being magnified[9,10]. However, problems of big data lie in the aspect of invasions of privacy and security. Big data are aggregations of numerous pieces of information of numerous individuals. Therefore, those who collect and analyze big data can become a figure of a big brother that collects and manage even individuals' private information[7,11,23]. In addition, if collected data are spilled due to security problems, socially big problems may be caused because the spilled data will contain almost all people's information[7,13–15].

In the present paper, a scheme to conduct safe authentication of contents by inserting attribution information at the beginning and end of contents when

* *Author for correspondence*

big data users safely download and install contents or download contents through application is proposed. The proposed scheme transmits attribution information together with the contents provided by the server to identify the identity of the server that provides the contents and the fact that the contents have not been altered so that the user who is being provided with the service can receive contents continuously without ceasing. The proposed scheme creates attribution information in the first and last contents to provide both authentication and non-repudiation even when one of two signatures has been lost thereby providing safety against malware attacks in order to provide contents services in real time.

This paper is composed as follows. In chapter 2, big data and malware attacks are examined. In chapter 3, a dual signature based contents protection scheme is proposed, in chapter 4, the efficiency of the proposed scheme is evaluated, and finally, in chapter 5, conclusions are drawn.

2. Ase of Use

2.1 Big Data

Big data refer to large scaled data which are much huger in scale compared to those data that have been created in analog environments in the past, are created in short cycles, and include not only numerical data but also letter data and video data[1,4,11]. Recently, as PC, the Internet, and mobile devices have become a way of life, data that are easily used and stored in cyber spaces without being restricted by time or places have been exponentially increasing. The explosive increase of digital information is also attributable to the expansion of Machine to Machine (M2M) communication that means information exchanges between humans and machines and between machines and machines[6,10,16].

Video contents including UCC created firsthand by users and letters created in mobile phones and SNS (Social Network Service) show different aspects from previous contents not only in data increase speed but also in the forms and quality. In particular, text information distributed in blogs or SNS enables the analysis of not only the propensity of the writers but also connections to the other parties of communication. The video information taken by CCTVs installed not only in major roads and public buildings but also even in apartment elevators is also stored as data. In addition, data are being mass-produced not only in the private sector but also in the public sector including diverse social surveys such as censuses, international data, health insurance, and pensions[8,17,18].

Big data are generally characterized by 3V, data volumes, data creation velocity, and the variety of forms. Although the diverse and huge data constituting big data are utilized as important resources that determine the competitive advantages of countries, the paradigm should be shifted in terms of not only the volume but also the quality and diversity of data which are much different compared to the past[5,11,19].

Big data enable analyzing large scaled customer information within a shorter time compared to the past utilizing technologies such as distributed processing. Big data users can grasp customer responses to their products and services by analyzing company related search words and comments created in Twitter and the Internet to immediately respond to the customer responses[16,20–22].

Big data also enable efficient system operation without constructing database-based expensive data warehouses because Hadoop in the form of open sources, analyzing package R, parallel analysis processing technologies, and cloud computing, etc. are utilized as software or hardware for big data.

2.2 Malware Attack

Malware refers to malicious codes including viruses, worms, and Trojan horses. Malware is largely divided into two types; methods that use communication tools and methods that find out vulnerability in systems[2,6,15]. Malware is spread using widely used communication tools including sending worms through e-mails or instant messages, circulating Trojan horses on websites, and connecting to P2P (Peer-to-Peer) to download files infected with viruses. Malware finds out existing vulnerability of systems to easily infiltrate into the systems before users know[15,17,18,22].

Malware acts secretly by hiding itself in system or not exposing its figure to users. Actions against malware are as follows.

- Download only the attached files of e-mails or instant messages sent by sources reliable or expected to be reliable.
- Inspect files attached to e-mails using Norton Internet Security before opening the files.
- Delete all undesired messages without opening them.

- Do not clink web links sent by unknown persons.
- Terminate the instant message session if a person in the friend list sent any strange message, file, or website link.
- Inspect all files using an Internet Security solution before sending them to your system.
- Transmit only those files that came from well-known sources.
- Block all unnecessary external communications using Norton Internet Security.
- Always maintain security patches in the newest state.

3. Smart Contents Protection Scheme using Fourier Series

In this section, a scheme that will enable users to be continuously provided with contents through the attribution information of the contents when smartphone users safely download and install contents or download contents through applications. The proposed scheme enables users to safely receive contents in real time without ceasing by inserting attribution information at the beginning and end of contents that users want to download.

3.1 Overview

To verify that contents provided by contents servers have not been altered, the proposed scheme transmits the contents provided by contents servers together with authentication information plus attribution information. The proposed scheme inserts attribution information at first and last contents so that all of integrity, authentication and non-repudiation can be provided for the contents even if one of the two signatures is lost thereby providing safety against malware attacks for provision of contents services by 3rd parties in real time.

Figure 1 shows the process of operation of the proposed scheme between contents servers and users. The authentication process between contents servers and users consists of four stages and contents data attribution information is transferred in stage 2 together with authentication parameters. As shown in Figure 1, each server has an authentication certificate, unique key and secret information used for electronic signatures/decryption. Users support both embedded memories and external memories of smartphones as well as the storage of secret information through USIM cards. The proposed scheme is assumed to support URL protocols, all platforms for communication with different applications or web browsers, and standard forms. This assumption is made because it is not necessary to automatically execute services in advance when a URL or protocol is called.

3.2 Formation of Attribution Information of Contents Data through Fourier Series

In this section, Fourier series are used to create attribution information on contents provided to smartphone users. Fourier series are used to determine whether data are

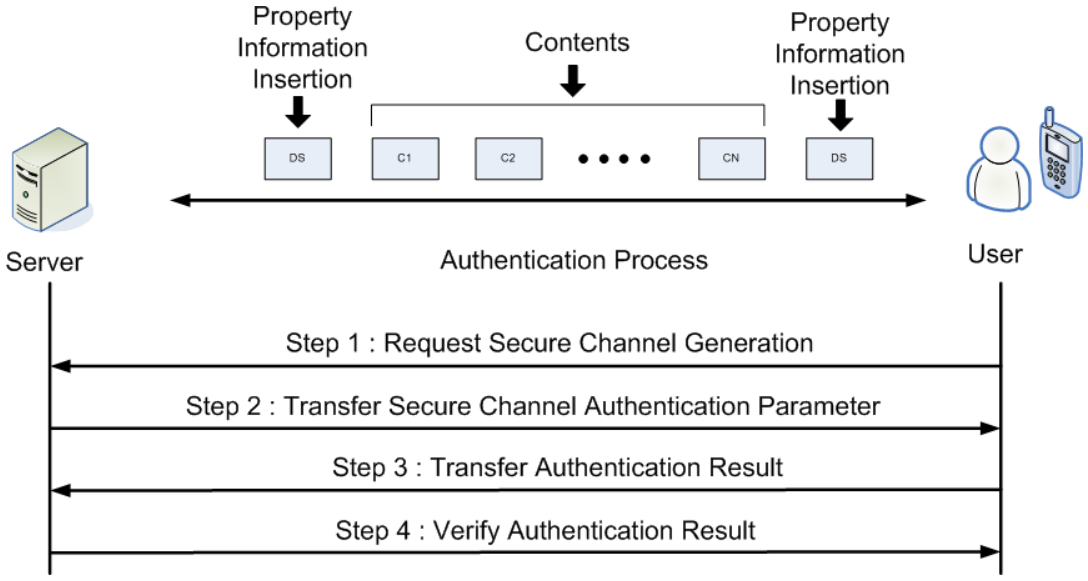

Figure 1. Processing environment of proposed scheme.

normal or not so that diverse contents provided to smartphone users can be served legitimately without ceasing in real time.

For all contents data *wt*, attribution information *f(wt)* satisfies the conditions set forth by expression (1).

$$f(wt)=\begin{cases} A\sin wt & (0\le wt\le\pi) \\ 0 & (\pi\le wt\le 2\pi) \end{cases} \quad (1)$$

Where, *A* refers to the size of attribution information, and the size of *A*'s ranges from 0 at the minimum to *A* at the maximum including *A*. Contents data are selected from between 0 and 2π.

Not only to serve the contents service provided by smartphone servers to users in real time but also to protect the contents service from 3rd parties, when smartphone users safely download and install contents or download the contents through applications, the proposed scheme develops attribution information function *f(wt)* using Fourier series to obtain coefficients a_0, a_n and b_n as follows.

$$a_0=\frac{1}{2\pi}\int_0^{2\pi} f(wt)d(wt)=\frac{1}{2\pi}\int_0^{\pi} A\sin wtd(wt) \quad (2)$$
$$=\frac{A}{2\pi}[-\cos wt]_0^{\pi}=\frac{A}{2\pi}(1-\cos\pi)=\frac{A}{\pi}$$

$$a_n=\frac{1}{\pi}\int_0^{2\pi} f(wt)\cos nwt\, d(wt)=\frac{1}{\pi}\int_0^{\pi} A\sin wt\cos nwt\, d(wt)$$
$$=\frac{A}{\pi}\int_0^{\pi}\frac{1}{2}\{\sin(n+1)wt-\sin(n-1)wt\}d(wt)$$
$$=\frac{A}{2\pi}\left\{\left[\frac{-\cos(n+1)wt}{n+1}+\frac{\cos(n-1)wt}{n-1}\right]_0^{\pi}\right\}$$
$$=\frac{A}{2\pi}\left\{-\frac{(-1)^{(n-1)}-1}{n+1}+\frac{(-1)^{(n-1)}-1}{n-1}\right\} \quad (3)$$
$$=\begin{cases} 0 & (n=1,3,5,\ldots) \\ \dfrac{-2A}{(n^2-1)\pi} & (n=2,4,6,\ldots) \end{cases} \quad (4)$$

$$f(wt)=\begin{cases} A\sin wt & (0\le wt\le\pi) \\ 0 & (\pi\le wt\le 2\pi) \end{cases}$$

$$=\begin{cases} \dfrac{A}{2} & (n=1) \\ 0 & (n\ne 1) \end{cases} \quad (5)$$

Attribution function *f(t)* is expressed as per expression (6).

$$f(t)=a_0+\sum_{n=1}^{\infty}a_n\cos nwt+\sum_{n=1}^{\infty}b_n\sin nwt \quad (6)$$

Using coefficients a_0, a_n, b_n obtained as shown above using expression (6), attribution function *f(t)* is obtained as shown under expression (7).

$$f(t)=\frac{A}{\pi}-\frac{2A}{\pi}\left(\frac{\cos 2wt}{3}+\frac{\cos 4wt}{15}+\frac{\cos 6wt}{35}+\ldots\right)+\frac{A}{2}\sin wt \quad (7)$$

In expression (7), as partial sums of attribution information increase, the square-wave original attribution information function *f(t)* is approached. That is, Fourier series converge and the sum becomes the given square-wave function *f(t)* of 2π which the cycle is the same as the fundamental wave. In addition, all partial sums at attribution information function *f(t)*'s discontinuous points x = 0, π, 2π,… become 0. The sums of series at attribution information function *f(t)*'s discontinuous points are the arithmetic means of the left and right limiting values of *f(t)* and the sums at points other than the foregoing points are *f(t)*.

$$f(t)=A\left\{\frac{1}{\pi}-\frac{2}{\pi}\left(\frac{\cos 2wt}{3}+\frac{\cos 4wt}{15}+\frac{\cos 6wt}{35}+\ldots\right)+\frac{1}{2}\sin wt\right\} \quad (8)$$

3.3 Contents Data Authentication Process

This process describes the process to authenticate the contents data transferred by contents servers to users. For users to serve contents services in real time, attribution information is inserted at the beginning and end of contents so that authentication and non-repudiation for the contents can be provided even when one of the two pieces of attribution information has been lost.

Through the key chain of the hash chain, the name node stores key K_i that is strictly shared with other nodes and set seed value $C_{j,o}$ to create and verify delegation tokens. Seed value $C_{j,o}$ is created when the hash function is applied *n* times to random numbers $C_{i,j}$. Where, $C_{i,j}$ refers to the random numbers of clients *i* and *j* . Clients are divided into many groups of at least *n* clients. This process is a process to conduct authentication between clients and data nodes. The authentication between clients and data nodes is composed of six stages as follows.

- Stage 1: The client encrypts (r_s, TK_s), t_s, and $C_{j,o}$ using secret key $K_{i,j}$ shared by the client and the data node and encrypts $K_i \,||\, K_j$ using seed value $C_{j,o}$. When all encryptions have been completed, the client transfers block access token (r_s, TK_s), time stamp t_s, and seed value $C_{j,o}$ to the data node as shown by expression (9).

$$Transfer \quad E_{K_{i,j}}((r_s,TK_s),t_s,c_{j,0}),K_i,E_{c_{j,0}}(K_i \,||\, K_j) \quad (9)$$

- Stage 2: After receiving the cipher text transferred from the client, the data node decrypts expression (10) using secret key $K_{i,j}$ shared by the client and the data node.

$$D_{K_{i,j}}((r_s,TK_s),t_s,c_{j,0}) \quad (10)$$

- Stage 3: The data node obtains seed value $C_{j,o}$ through

expression (10). Using seed value $C_{j,o}$ and key K_i, the data node decrypts cipher text $E_{c_{j,0}}(K_i \,||\, K_j)$ as shown by expression (11) to obtain K_j key K_j.

$$D_{K_{i,j}}(K_i \,||\, K_j) \quad (11)$$

- Stage 4: After identifying time stamp t_s, the data node authenticates block access token (r_s, TK_s). If the user is a normal user, the data node encrypts the data using secret key $K_{i,j}$ and transfers the encrypted data to the client.

$$\text{Check } (r_s, TK_s), t_s \quad (12)$$

$$\textit{Transfer} \quad E_{K_{i,j}}(Data \,||\, f(t)), E_{c_{j,0}}(f(t)) \quad (13)$$

- Stage 5: After receiving expression (13), the data node decrypts the data using secret key $K_{i,j}$. If the user is a normal user, the data node regenerates $E_{K_{i,j}}((r_S, TK_S), t_S, c_{j,0}), K_i, E_{c_{j,0}}(K_i \,||\, K_j)$ and send them to the client as shown by expression (14).

$$\text{Regenerate} \quad E_{K_{i,j}}((r_S, TK_S), t_S, c_{j,0}), K_i, E_{c_{j,0}}(K_i \,||\, K_j) \quad (14)$$

- Stage 6: The client decrypts the data transferred by the data node using secret key $K_{i,j}$. The client checks the decrypted data to see if they are normal data using attribution information function $f(t)$.

4. Evaluation

4.1 Experimental Environment

To maintain the objectivity of the proposed scheme, an experimental environment was constructed referring to the model set forth under[10]. The total number of packets was set to 10,000 and the experiment was conducted so that packets sized 64, 128, 256, and 512 bytes were transmitted using UDP. The queue size of the transmitter/receiver was set to 100 and the processing rate of processing packets per second was set to 10. Experiments were conducted for each of packet drop rates 10%, 20%, 30%, 40%, 50%, and 60% and the average length of burst drops was set to 10.

Table 1. Simulation parameter

Parameter	Value
Total packet number	10,000
Packet size	64, 128, 256, 512
Queue size	100
Packet process rate	10
Packet drop rate	10%, 20%, 30%, 40%, 50%, 60%
Average Length of Burst drop	10

4.2 Results and Analysis

Figure 2 shows the evaluation of delay times according to packet loss rates of contents when a smart user has requested the server for contents services. The proposed scheme has an advantages of being applicable to applications in real time because the receiver can authenticate the transmitter on receipt of the first packet and another advantage of being able to receiver's processing time because non-repudiation is provided even if the signature included in the last packet is not verified unless the first packet is lost during transmission.

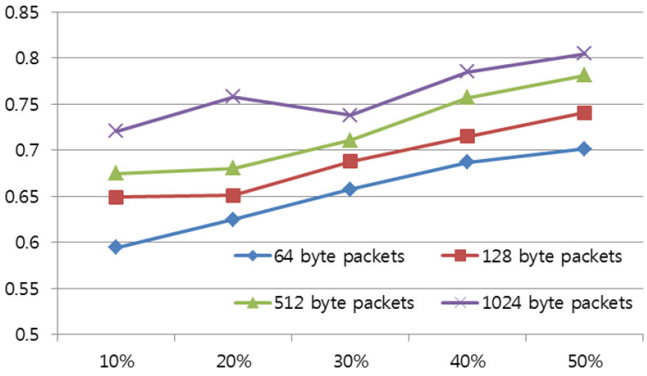

Figure 2. Delay time 9ms/per packet loss rate.

Figure 3 shows a comparison and evaluation of throughput of contents processed per packet. As shown by the results in Figure 3, contents throughput increased in proportion to packet loss rates. Since authentication information for contents not immediately neighboring is transmitted when packet losses are small, authentication success rates show differences by up to 3.8% according to differences in packet loss rates.

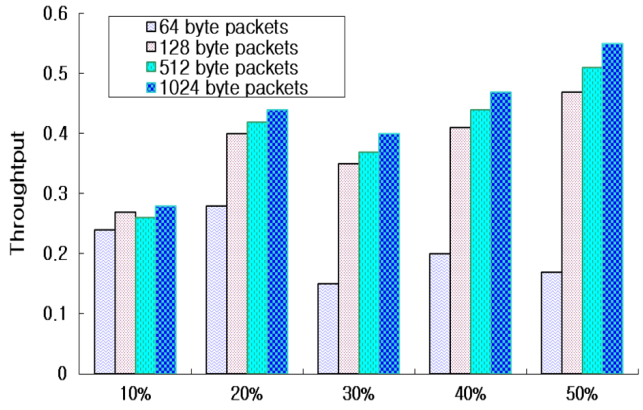

Figure 3. Throughput per packet loss rate.

4.3 Security Analysis

In the proposed scheme, user privacy information is generated every session between the user and the smartphone using the random number generated by the individual user and the random number generated by the user belonging to the service group. Since user privacy information is generated every time the user accesses smartphone services in the proposed scheme, disturbance by 3rd parties is prevented while the individual user is reading the privacy information of the user belonging to the service group.

In the proposed scheme, since the random number generated by a user cannot be known by other users and user privacy information is generated using new random number every session, data are safe against location tracking attacks. The privacy information generated by the user belonging to the service group makes 3rd parties' verification whether the user's response in the current session is the same as the response tapped in previous sessions difficult.

In the proposed scheme, to protect smartphone users' privacy, each user registers his/her information in the server before being provided with smartphone services. When user registration in the server has been completed, the user can safely protect his/her privacy information from 3rd parties by determining whether any 3rd part is accessing the service using the user's privacy information. The user privacy information in the smartphone and the security recognizer makes it impossible for 3rd parties to guess the random number generated by the user of the smartphone and the random number generated by the user belonging to the service group. Therefore, the proposed protocol provides the anonymity of users.

5. Prepare your Paper before Styling

In the present paper, a scheme was proposed that will include authentication information in contents when smartphone users safely download and install contents or download contents through applications to enable users to continuously use contents. The proposed scheme inserted electronic signatures at the beginning and end of contents so that users can download contents from servers without ceasing in real time safely against malware attacks. In future studies, the results of the present study will be actually applied to smartphones.

6. Acknowledgment

This work was supported by the Security Engineering Research Center, granted by the Ministry of Trade, Industry and Energy.

7. References

1. Becher M, Freiling C, Hoffmann J, Holz T, Uellenbeck S, Wolf C. Mobile security catching up? Revealing the nuts and bolts of the security of mobile devices. Proceedings of IEEE Symposium on Security and Privacy; 2011. p. 96-111.
2. Zhou W, Zhou Y, Jiang X, and Ning P. Detecting repackaged smartphone applications in third-party android marketplaces. Proceedings of the 2nd ACM Conference on Data and Application Security and Privacy; 2012. p. 317-326.
3. Manyika J, Chui M. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute; 2011; p. 1.
4. Gantz J, Reinsel D. Extracting Value from Chaos. IDC IVIEW. 2011; 6.
5. Jung YC. Big data revolution and media policy issues. KISDI Premium Report. 2012; 12(2):1–22.
6. Sahs J, Khan L. A machine learning approach to android malware detection. IEEE Intelligence and Security Informatics Conference (EISIC) 2012 European; 2012. p. 141–7.
7. Kim JT, Oh BJ, Park JY. Standard trends for the bigdata technologies. Electronics and Telecommunications Trends. 2013; 28(1):92–9.
8. Lee SH, Lee DW. Current status of big data utilization. Journal of Digital Convergence. 2013; 11(2):229–33.
9. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH. Big data: The next frontier for innovation, competition and productivity. Mckinsey Global Institute. 2011; p. 1–137.
10. Barakat OL, Hashim SJ, Raja Abdullah RSAB, Ramli AR, Hashim F, Samsudin K, Rahman MA. Malware analysis performance enhancement using cloud computing. Journal of Computer Virology and Hacking Techniques. 2014; 10(1):1–10.
11. Vatamanu C, Gavilut D, Benchea RM. Building a practical and reliable classifier for malware detection. Journal of Computer Virology and Hacking Techniques. 2013; 9(4):205–14.
12. Jeong YS, Han KH. Service management scheme using security identification information adopt to big data environment. Journal of Digital Convergence. 2013; 11(12):393–9.
13. Tankard C. Big data security. Network Security. 2012; p. 5–8.
14. Cao Y, Miao Q, Liu J, Gao L. Abstracting minimal security-relevant behaviors for malware analysis. Journal of Computer Virology and Hacking Techniques. 2013; 9(4):171–8.
15. Hong JK. Kerberos Authentication Deployment Policy of US in Big data Environment. Journal of Digital Convergence. 2013; 11(11):435–41.

16. Cimpoesu M, Gavilut D, Popescu A. The proactivity of perceptron derived algorithms in malwawre detection. Journal of Computer Virology and Hacking Techniques. 2012; 8(4):133–40.
17. Bose A, Hu X, Shin K, Park T. Behavioral detection of malware on mobile handsets. Proceedings of the 6th International Conference on Mobile Systems Applications and Services; ACM; 2008. p. 225–38.
18. Amrukar C, Traynor P, Oorschot P. Measuring ssl indicators on mobile browsers: Extended life, or end of the road? Information Security. 2012; p. 86–103.
19. Russom P. Big Data Analytics. TDWI Research Fourth Quarter. 2011; 6.
20. Son SY. Big data, online marketing and privacy protection. KISDI Premium Report. 2013; 13(1):1–26.
21. CDWG. Proactive planning for big data. CDWG–People Who Get It. 2013; p. 1–8.
22. Hong JK. The security policy for big data of US government. Journal of Digital Convergence. 2013; 11(10):403–9.
23. Lane A. Securing big data: Security recommendations for hadoop and nosql environments. Securosis. 2012; 1-16.