Scientific Research Publishing

Communications and Network, 2017, 9, 291-301
http://www.scirp.org/journal/cn
ISSN Online: 1947-3826
ISSN Print: 1949-2421

# Towards the Development of Best Data Security for Big Data

**Yuan Tian**

Department of Computer Science, King Saud University, Riyadh, Saudi Arabia
Email: ytian@ksu.edu.sa

**How to cite this paper:** Tian, Y. (2017) Towards the Development of Best Data Security for Big Data. *Communications and Network*, **9**, 291-301.
https://doi.org/10.4236/cn.2017.94020

**Received:** November 2, 2017
**Accepted:** November 27, 2017
**Published:** November 30, 2017

Copyright © 2017 by author and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

## Abstract

Big data is becoming a well-known buzzword and in active use in many areas. Because of the velocity, variety, and volume of big data, security and privacy issues are magnified, which results in the traditional protection mechanisms for structured small scale data are inadequate for big data. Sensitivities around big data security and privacy are a hurdle that organizations need to overcome. In this paper, we review the current data security in big data and analysis its feasibilities and obstacles. Besides, we also introduced intelligent analytics to enhance security with the proposed security intelligence model. This research aims to summarize, organize and classify the information available in the literature to identify any gaps in current research and suggest areas for scholars and security researchers for further investigation.

## Keywords

Big Data Analytics, Secure Big Data, Security Intelligence Model

## 1. Introduction

Big data is transforming the center of modern science and global business landscape. These data are generated from multitude of devices, online transactions, health records, search queries, social network-related information, videos, audios, images, and etc. The proliferation of these data is created with incredible volume, velocity, and variety, which is known as 3 Vs [1] [2] [3] (shown in **Figure 1**). The lifecycle of big data is the process of analyzing large amounts of data of a variety of types to produce unknown correlations and hidden information.

The first "V" represents the volume of data. In the past decade the size of data increases exponentially and nowadays it's getting very common for enterprises to have storage system more than terabytes or even petabytes. Next come to the

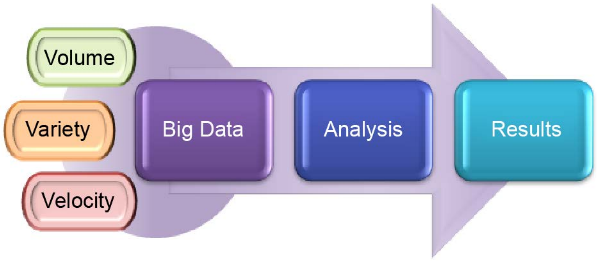

Figure 1. The lifecycle of big data.

variety of data. The types of data vary in different ways rather than just text data. Sometimes the data is not in a traditional format or even in a form that we have not thought about it yet. Generally the types of data can be divided into three categories: structured, semi structured, and unstructured. Structured data refers to well organized and easily sorted data, while unstructured data represents data which is random and hard to analyze. Semi-structured data does not fix to a certain field but uses tags to differentiate data elements. Last of the three "V"s is the velocity of data. It refers the speed of the data that we discussed above. Every day billions of messages are generated from social media like Facebook or Tweeter. Huge amount of data is produced in no time and next second these messages (status or a tweet) are not new and interesting to users as they only focus on the recent updates. The data movement is now almost real time and the update window has reduced to fractions of the seconds.

The traditional data ware house comes under the structured type of data and hence big data in no way eliminates the need for traditional data warehousing, but just includes it in a bigger data set and takes it to the next level. The data in Figure 2 illustrates the status of vulnerability and threats in cyber security over the last 12 months in 2013. We can see that the rate and complexity of cyber security grows continually. Consequently, organizations need to make a quickly move to prevent costly and brand damaging security incident from happening.

One of the key security concerns related to big data analysis and aggregation is that a huge amount of sensitive data of individuals is collected and examined by organizations. In order to gain values from information like trade secrets, financial records, or intellectual properties, organizations are increasingly collecting such data from stores and applications. [4]

## 2. Current Data Security for Big Data

- Authentication

Authentication is the process to determine whether the identity of the users, services, and hosts are whom they claim to be. The process of identifying an entity usually based on user name and password. In computing systems, authentication is differ from authorization, the former one merely exams that the entity is who he claims to be but not giving access rights to system objects based on their identity. Authentication and authorization must work in tandem to provide effective security.

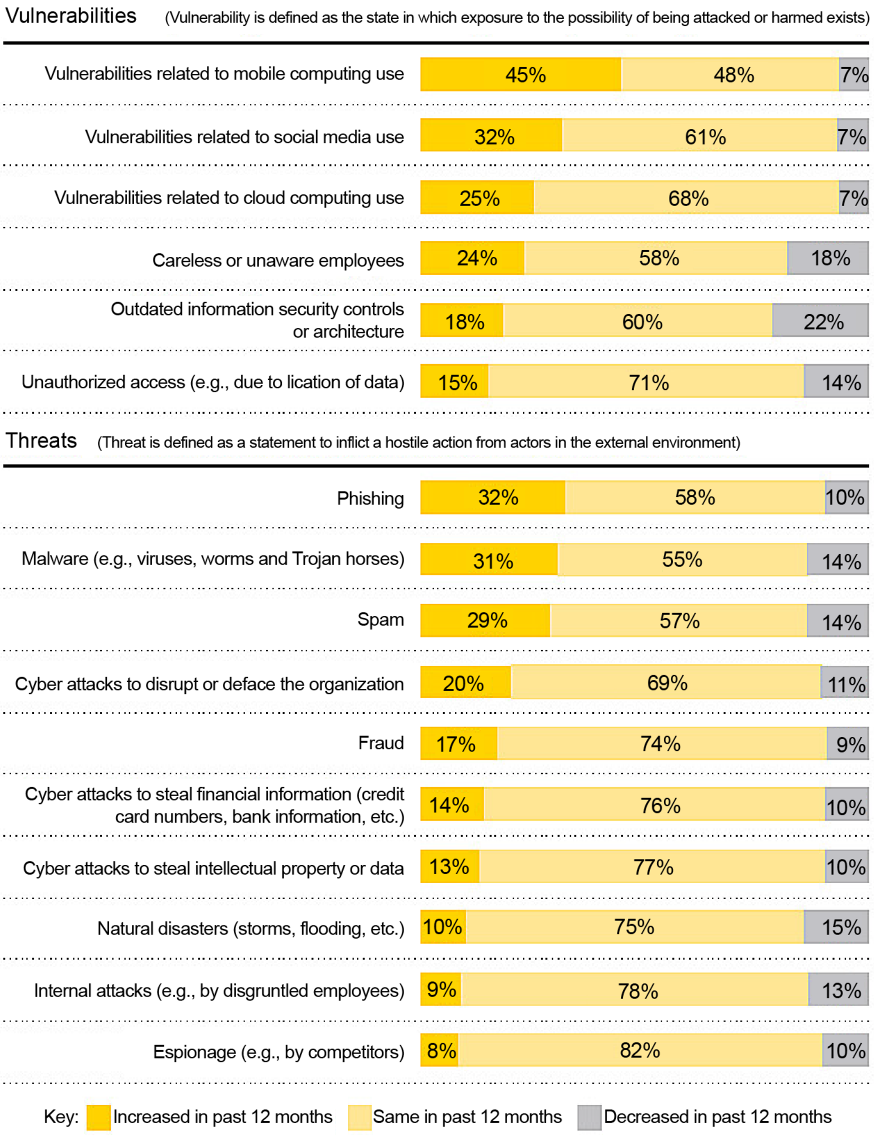

Figure 2. The trend of vulnerability and threats in cyber security [5].

- Authorization

Authorization is the process to determine which permissions a person, data, service, system is supposed to have. In multi-user computing systems, a system administrator defines which users are allowed access to the system, as well as the privileges of use for which they are eligible (e.g., access to file directories, hours of access, amount of allocated storage space). Authorization can be seen as both the preliminary setting of permissions by a system administrator, and the actual checking of the permission values when a user obtains access. Authorization is usually preceded by authentication.

- Data Protection

Ensure that only authorized users have access to accurate and complete information when required. The main goal is to guarantee data is appropriately protected from modification or disclosure.

- Auditing

Security auditing is a manual or systematic measurable technical assessment of a system or application, which ensures a permanent record about who did what at which time. Manual assessments include interviewing staff, performing security vulnerability scans, reviewing application and operating system access controls, and analyzing physical access to the systems.

Security issues have to be solved in order to capture the full potential of big data. Consequently, related security policies need to be applied to big data world. Besides, organizations should also put the right technologies in place and structure workflows and incentives for the best usage of big data. As a result, organizations need a way to protect, utilize, and gain real-time insight to achieve secure big data.

In this work, we give an overview of big data and show the competitiveness it has. We designed a novel security model for big data to meet the security challenges. The remainder of this paper is organized as follows. Section 2 discusses the feasibility and obstacles in the implementation process of big data. In Section 3, we propose a secure intelligence model for achieving secure big data. Section 4 summarizes the exiting literatures regarding security and privacy protection methods in big data. Finally, the conclusion is presented in Section 5.

## 3. Feasibility and Obstacles of Big Data

### 3.1. Competitiveness and Value of Big Data

The use of big data is becoming a key basis of competition for companies and leading individual firms to outperform their competitors. Forward-thinking leaders should begin aggressively to build their organizations' big data capabilities from the standpoint of innovation, competition, and capture of value. The collection of data and analysis can lead companies for better performance output and making better management decisions.

Indeed, early adopters for using big data are found in many scenarios. For example, transactional data are created and stored in digital form in organizations for more accurate, variable, and improved performance output, which ranges from product inventories to emergency leave days. In healthcare domain, physicians and researchers examine the illness status if a certain medicine has been wildly prescribed. Pros and cons can by observed from the healthcare outcomes. Other data pioneers collect data from sensors to help to design future products. They embed sensors into children's toys in order to see how their products are actually used in the real world.

### 3.2. Challenge Issues for Achieving Secure Big Data

Although the applications of big data are advanced in many aspects, we must address several security challenges to realize its true potential. The Cloud Security Alliance (CSA) [6] established a big data working group in 2012. The latest report they issued discussed the top security and privacy issues for big data. This report details top ten big data specific security and privacy challenges in order to bring renewed focus on reinforcing the infrastructure of big data. Table 1 summarizes the main obstacles and countermeasures for the growth of big data as discussed in [6] [7] [8].

Despite of these obstacles as well as opportunities and advantages, cloud computing raises several security issues and hence security is still the primary concern of many customers who want to leverage public cloud services.

Table 1. Top ten big data security challenges.

| 1. Security Types | 2. Scenario | 3. Challenges | 4. Description | 5. Threats | 6. Current Mitigations |
|---|---|---|---|---|---|
| Infrastructure Security | Application Computation Infrastructure | Secure computations in distributed programming frameworks | Distributed programming frameworks utilize parallelism in computation and storage to process massive amounts of data. | Malfunctioning compute worker nodes<br>Access to sensitive data<br>Privacy of output information | Trust establishment: initiation, periodic trust update<br>Mandatory access control<br>Privacy preserving transformations |
| | Data from Diverse Appliances and Sensors | Security best practices for non-relational stores | Non-relational data stores popularized by NoSQL databases are still evolving with respect to security infrastructure. | Lack of stringent authentication and authorization mechanisms<br>Lack of secure communication between compute nodes | Enforcement through middleware layer<br>Passwords should never be held in clear<br>Encrypted data at rest<br>Protect communication using SSL/TLS |
| Data Management | Consumer Data Archive | Secure data storage and transactions logs | The exponential increasing of data set requires auto-tiering for big data storage management. | Data Confidentiality and Integrity<br>Availability<br>Consistency<br>Collusion | Encryption and signatures<br>Proof of data possession<br>Periodic audit and hash chains<br>Policy based encryption |
| Data Management | Audit of usage, pricing, billing | Granular audits | In order to be notified at the attack takes place, we need audit information. | Completeness of audit information<br>Timely access to audit information<br>Integrity of audit information<br>Authorized access to audit information | Infrastructure solutions as discussed before. Scaling of SIEM tools. |
| | Keeping track of ownership of data pricing, audit | Data provenance | Analysis of large provenance graphs to detect metadata dependencies for security/confidentiality applications is computationally intensive. | Secure collection of data<br>Consistency of data and metadata<br>Insider threats | Authentication techniques<br>Message digests<br>Access Control through systems and cryptography |

**Continued**

| | | | | | |
|---|---|---|---|---|---|
| Ingegrity and Reactive Security | Data Poisoning | End-point input validation/ filtering | Big data technology can provide fast processing and various types of data analysis. | Adversary may tamper with device or software | Tamper-proof software |
| | | | | Adversary may clone fake devices | Trust certificate and truste devices |
| | | | | Adversary may directly control source of data | Analytics to detect outliers |
| | | | | Adversary may compromise data in transmission | Cryptographic Protocols |
| | Fraud Detection | Real time security compliance monitoring | Detecting in a real-time manner for anomalous retrieval of personal information. | Security of the infrastructure | Discussed before |
| | | | | Security of the monitoring code itself | Secure coding practices |
| | | | | Security of the input sources | Discussed before |
| | | | | Adversary may cause data poisoning | Analytics to detect outliers |
| Data Privacy | Consumer Data Privacy | Scalable and composable privacy preserving data mining and analytics | User safety will be improved if scalable and robust privacy preserving data mining algorithm are applied. | Exploiting vulnerability at host | Encryption of data at rest, access control and authorization mechanisms |
| | | | | Insider threat | Separation of duty principles, clear policy for logging access to datasets |
| | | | | Outsourcing analytics to untrusted partners | Unintended leakage through sharing of data |
| | | | | Unintended leakage through sharing of data | |
| | Data Integrity and Privacy | Cryptographically enforced access control and secure communication | To ensure that the most sensitive private data is end-to-end secure and only accessible to the authorized entities, data has to be encrypted based on access control policies. | Enforcing access control | Identity and Attribute-based encryptions |
| | | | | Search and filter | Encryption techniques supporting search and filter |
| | | | | Outsourcing of computation | Fully Homomorphic Encryption |
| | | | | Integrity of data and preservation of anonymity | Group signatures with trusted third parties |
| | Data Privacy | Granular access control | The shared data is often swept into a more restrictive category to guarantee sound security. | Keeping track of secrecy requirements of individual data elements | Pick right level of granularity: row level, column level, cell level |
| | | | | Maintaining access labels across analytical transformations | At the minimum, conform to lattice of access restrictions. More sophisticated data transforms are being considered in active research |

### 3.3. Achieving Best Data Security for Big Data

- Massively Scalable Data Security

Granularly control over users to determine who can store, access and process massive, dynamic, and potentially sensitive data.

- Maximum Transparency

Information transparent helps big data unlock significant value at much higher frequency. On the on hand, organization should find insights in the data assets and create value based on the observation. On the other hand, the use of these data should keep transparent. There is a great need to build trust and transparency framework and mechanism around personal data to guarantee that individual's privacy preferences are considered. Thus, companies should mature their governance polices by offering users full transparence control of the way their personal data is used.

- Maximum Performance

Variety of best practices for optimization techniques in big data and maximizing application performance are required. For example, Apache Hadoop software [9] provides faster performance in data acquisition phase and same or better performance in the extraction and analysis phase.

- Easy to Use

In a highly technical environment, data experts play an important role as organizations need such skilled person to mine data for insights and make decisions act on them. Obviously, having more data scientists are great, but an alternative solution is to create analytics products which are easy to user even for common people. Products should provide easy to use software applications for fast analytics and visualization and the goal is to help people see and understand data without difficulty.

- Heterogeneous System Compatibility

Big data aims to compatible to heterogeneous system although it may perform differently in terms of performance, reliability or some other characteristics. Besides, compatible use of big data requires that the collection of personal data should always compatible with further processing.

- Enterprise Ready

The adoption of piloting big data into large enterprises expends quickly for the reason that companies need big data to analyze the internal information flows. Situations like fraud detection, network maintenance, and customer service are all touched. A data fabric with Hadoop, analytics and data warehousing information are required by enterprises and many of those methods are already implemented.

## 4. The Proposed Security Intelligence Model for Big Data

The application of big data analytics to security issues and security protection for big data are two sides of the same coin. We need to improve big data security by not only applying traditional security mechanisms, but also introducing intelligent analytics to enhance security.

### 4.1. The Proposed Security Intelligence Model

In Figure 3, we describe a security intelligence model to achieve best security for big data. Various types of data are generated from diverse sources, such as application data, mobile-based data, and etc. The sources may number in hundreds and formats of data can be dozens.

These data generally is divided into two categories, namely passive data (as known as long term data) and active data (also called real time data). Each category contains traditional structured data, which fits neatly into rows and columns and non-traditional semi or unstructured data. The detailed introduction of passive and active data is discussed in the following sub-section. The data then can be ingested by numerous tools. For example, use ETL to extract, transform and load data, or use Flume to stream log collection, or use Sqoop to transfer data between relational database and Hadoop, and so on.

Big data analytics is expected to spur changes in information security. The proposed intelligence analytics platform is capable of massive and diverse real time data collection and threat analysis. Security management driven by big data analysis creates a unified view of multiple data sources and centralizes threat research capabilities. It ingests external threat intelligence and also offers the flexibility to integrate security data from existing technologies. The platform

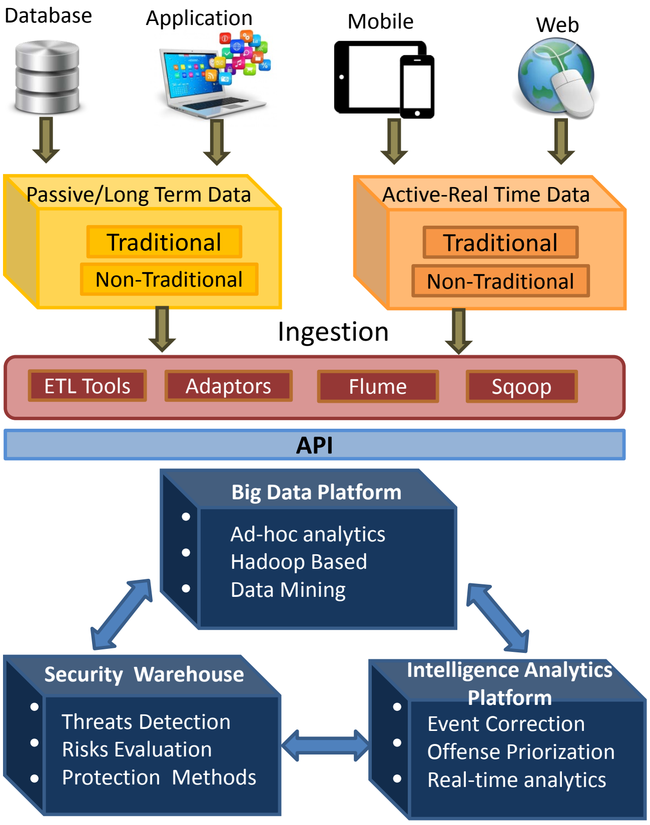

Figure 3. The proposed intelligence driven security model for big data.

provides event correction, offence priorization and real time analytics to gain insights of security intrusion. The intelligence analytics platform mines meaningful security information to protect big data.

Security warehouse stores the current protection methods, which used to tackle any security problems results from big data platform and intelligence analytics platform. It provides threats detection before security intrusion happens and evaluate the hazards of each risk. Intelligence analytics platform enhances the protection methods provided by security warehouse as it can analyze and correlate broader data sets to prevent cyber-attacks, physical threats, fraudulent claims, and account takeovers

### 4.2. Data Sources for Security Analytics

The source of data we discussed in **Figure 3** is divided into active and passive based on the summarization in [10]. Passive inputs include:

- Data generated from computer, for example, geographical IP location, E-health certificates, keyboard typing pattern reorganization, and click stream patterns.
- Data from Mobile: e.g., GPS location, network location, wireless access point.
- Data from physical access of user, which includes physical, assess time and location to the network.
- Data from human resource. For example, what is the user's role and what is this user's privilege if taking this role.
- Data from travel system. Traits of the travel contain source, destinations, and itineraries.
- Security information and event management (SIEM) data. SIEM systems collect access, logs and other security-related documentation like internal threats for analysis.
- Data from external sources, which includes IP blacklist and external threats.

Active input contains real time data sources, which include:

- Login information, such as user name and password.
- One-time passwords, which can be implemented by mobile phone, proprietary tokens, and text messaging.
- Digital identity certificates.
- Security questions or knowledge-based questions, in which the user is asked to answer at least one personal question.
- Biometric information, which is used to confirm the identity and determine the access profile of a person, such as a palm/finger print, face/voice recognition, and DNA.
- Social network data from Twitter, Facebook, Instagram, and etc.

## 5. Review Methodology

We have studied research works which related to security and privacy threats in big data. The summarized review in **Table 2** has been accomplished by reviewing

Table 2. Current security and privacy works in big data.

| Ref. | Context of Reseach | Problem Discussed | Technique Used | Model/Tool Proposed |
|---|---|---|---|---|
| [11] | Security for big data computing | Protect security of G-Hadoop system | Security model for the G-Hadoop framework | Yes |
| [12] | Big data security | Protecting for essential attribute of big data | Security hardening methodology with attributes relation graph | Yes |
| [13] | Protect value of big data | Prioritizesbattributes for big data security | Attribute selection methodology | Yes |
| [14] | Access control for big data | Content-centric information sharing of big data | Content-based access control model | No |
| [15] | Social issues of big data and cloud | Privacy and confidentially | Theory | No |
| [16] | Big data privacy preservation | Scalability issue of multidimensional anonymization over big data on cloud | A scalable multidimensional anonymization approach | Yes |

the exiting literatures regarding security and privacy issues in big data. The goal is to identify the current security and privacy protection methods, categorize them and suggest readers for further investigation

## 6. Conclusions and Research Indications

Big Data will help to create new growth opportunities and entirely new categories of companies, such as those that aggregate and analyze industry data. Meanwhile, with the momentum behind big data growing, a comprehensive security mechanism is needed to mitigate risk of breach and assure the best usage of big data technology. In this paper, an overview of big data and its related security issues are discussed. We also propose an intelligent security model for achieving best big data security.

## References

[1] Singh, S. and Singh, N. (2011) Big Data Analytics. 2012 *International Conference on Communication, Information & Computing Technology Mumbai India*, IEEE, October 2011.

[2] Gerhardt, B., Griffin, K. and Klemann, R. (2012) Unlocking Value in the Fragmented World of Big Data Analytics. Cisco Internet Business Solutions Group. http://www.unleashingit.com/docs/W13/Information-Infomediaries.pdf

[3] Sagiroglu, S. and Sinanc, D. (2013) Big Data: A Review, Collaboration Technologies and Systems (CTS). 2013 *International Conference on Digital Object Identifier*, 42-47.

[4] Tankard, C. (2012) Big Data Security. *Network Security*, **2012**, 5-8. https://doi.org/10.1016/S1353-4858(12)70063-6

[5] EY's Global Information Security Survey—2016, under Cyber Attack. http://www.ey.com/gl/en/services/advisory/ey-cybersecurity

[6] The Expanded Top Ten Big Dta Secrity & Privacy Challenges. https://cloudsecurityalliance.org/download/expanded-top-ten-big-data-security-an

d-privacy-challenges/

[7] Top Ten Big Data Security and Privacy Challenges, Cloud Security Allience, November 2012.

[8] Roy, A. (2013) Top Ten Security and Privacy Challenges for Big Data and Smartgrids, Fujitsu Laboratories of America.

[9] Allene, B. and Righini, M. Intel Distribution for Apache Hadoop Software, Better Performance for Big Data. https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/big-data-financial-services-better-performance-big-data-whitepaper.pdf

[10] Curry, S., Kirda, E., Schwarts, E., Stewart, W.H. and Yoran, A. (2013) Big Data Fuels Intelligence-Driven Security, RSA Security Brief, Jan. 2013.

[11] Zhao, J., Wang, L., Tao, J., Chen, J., Sun, W., Ranjan, R., Kołodziej, J., Streit, A. and Georgakopoulos, D. (2014) A Security Framework in G-Hadoop for Big Data Computing across Distributed Cloud Data Centres. *Journal of Computer and System Sciences*, **80**, 994-1007. https://doi.org/10.1016/j.jcss.2014.02.006

[12] Kim, S.-H., Eom, J.-H. and Chung, T.-M. (2013) Big Data Security Hardening Methodology using Attributes Relationship Information Science and Applications. *International Conference on Digital Object Identifier*, 1-2.

[13] Kim, S.-H., Kim, N.-U. and Chung, T.-M. (2013) Attribute Relationship Evaluation Methodology for Big Data Security, IT Convergence and Security. *International Conference on Digital Object Identifier*, 1-4.

[14] Zeng, W., Yang, Y. and Luo, B. (2013) Access Control for Big Data using Data Content. *IEEE International Conference on Big Data*, 45-47.

[15] Hayashi, K. (2013) Social Issues of Big Data and Cloud: Privacy, Confidentiality, and Public Utility. *International Conference on Availability, Reliability and Security*, 506-511. https://doi.org/10.1109/ARES.2013.66

[16] Zhang, X., Yang, C., Nepal, S., Liu, C., Dou, W. and Chen, J. (2013) A Map Reduce Based Approach of Scalable Multidimensional Anonymization for Big Data Privacy Preservation on Cloud. *International Conference on Cloud and Green Computing*, 105-112. https://doi.org/10.1109/CGC.2013.24