Big Data Security and Privacy: A Review

MATTURDI Bardi[1], ZHOU Xianwei[2], LI Shuai[2], LIN Fuhong[2*]

[1] School of Mathematics and Information, Hotan Teachers College, Hetian 848000, Xinjiang, P. R. China

[2] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, P. R. China

Abstract: While Big Data gradually become a hot topic of research and business and has been everywhere used in many industries, Big Data security and privacy has been increasingly concerned. However, there is an obvious contradiction between Big Data security and privacy and the widespread use of Big Data. In this paper, we firstly reviewed the enormous benefits and challenges of security and privacy in Big Data. Then, we present some possible methods and techniques to ensure Big Data security and privacy.

Keywords: big data; value of big data; security and privacy

I. Introduction

Recently, there has been an increasing interest in Big Data. However, the term Big Data remains vague. In Wikipedia, Big Data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications. A widely recognized definition belongs to IDC: “big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and/or analysis” [1].

We live in the Age of Big Data” [2]. In the past few years, the total amount of data created by human has exploded [3]. From 2005 to 2020, the amount of data is predicted to increase 300 times, from 130 exabytes to 40,000 exabytes [4]. These data are generated from scientific research, finance and business informatics, government, Internet search, social networks, document, photography, audio, video, logs, click streams, mobile phones, sensor networks and so on. Big Data is the result of the dramatic increase of data.

Big Data have significant value. Many organizations worldwide collect and analyze their own business process data in order to improve their internal decision making. The Obama regime has announced a Big Data research and development initiative based on recognition of the great social and economic value captured in the data in 2012 [5]. Big Data promotes the economy and scientific research, transforms traditional business models and scientific methods and creates new opportunities through data analysis [6].

However, exploring and using the extraordinary value of Big Data must increase risks of security and privacy. For example, “Amazon monitors our shopping preferences and Google learns our browsing habits, while Twitter knows what’s on our minds. Facebook seems to catch all that information too, along with our social relationships. Mobile operators know not only whom we talk to, but who is nearby. With Big Data promising valuable insights to those who analyse it, all signs seem to point to a further surge in others’ gathering, storing, and reusing our personal data.” [7]. If the Internet age threatened security and privacy, the age of Big Data endanger them even more.

Big Data security usually is to the use of the Big Data to implement solutions increasing security, reliability, and safety of a distributed system. Big Data privacy focuses on the protection of Big Data from unauthorized use and unwanted inference [8].

It is well known that big data are a priceless source of information at the basis of robust and accurate security solutions. However, Big Data often contain sensitive information that needs to be protected from unauthorized access and release. Obviously, there are not any challenges of security and privacy if we do not extract value from Big Data. Thereby, the principles of Big Data security and privacy must be balanced against additional societal value of Big Data.

In this paper, we firstly reviewed the enormous benefits and challenges of security and privacy in Big Data. Then, we present some possible methods and techniques to ensure security and privacy in Big Data. The rest of paper is as follows: section II shows some detail of Big Data, including definition, characteristic, framework, technology, example, value, and challenges of Big Data; section III reveals Big Data security and privacy issues and some solutions of Big Data security and privacy; section IV is the conclusion of this paper.

II. Big Data

2.1 Definition and characteristic of Big Data

Big Data become a popular topic in many fields such as scientific research, finance and business since McKinsey & Company, a global consulting agency announced the report “Big Data: the next frontier for innovation, competition, and productivity” in May 2011[9]. Today, people still have different opinions on its definition although the importance of Big Data has been generally recognized. Research scholars, data analysts and technical practitioners have different definitions of Big Data as different concerns. In Wikipedia, “big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications”. In the report of McKinsey & Company, “big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze”. In both definitions, datasets can be considered as Big Data will grow as technology advances over time and can vary by sectors. It can be seen that the volume of a data is not the only criterion for Big Data. Apart from masses of data, Big Data has some other features which determine the difference between itself and massive data.

In a 2001 research report, challenges and opportunities brought by increased data are defined as a 3Vs model, i.e., Variety, Velocity and Volume [10]. Gartner and many other enterprises used the “3Vs” model to describe Big Data [11]. In 2012, Gartner updated this definition: “Big Data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.”[12] In the “3Vs” model, Variety indicates the various types of data which include structured, semi-structured and unstructured data; Volume means data scale is large; Velocity implies all processes of Big Data must be rapid and timely in order to maximize value of Big Data. These features that Big Data handles large amount of data and utilizes various types of data including unstructured data and attributes that were never used in the past distinguish data mining from Big Data.

In 2011, IDC defined big data as “Big Data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and/or analysis” [1]. In this definition, features of big data may be summarized as 4Vs, i.e., Variety, Velocity, Volume and Value, where the implications of Variety, Velocity, Volume is identical the 3Vs model respectively and Value refers big data have great social value. The 4Vs model was widely recognized because it indicates the most critical problem which is how to discover value from an enormous, various types, and rapidly generated datasets in big data.

2.2 Example and value of Big Data

Many organizations worldwide have carried out big data action because they realize that the immense social and economic value is explored and used from big data.

Science research such as the Large Hadron Collider experiments [13], the Square Kilometre Array telescope [14], the Sloan Digital Sky Survey and the Large Synoptic Survey Telescope [3], the NASA [15] is producing or will create vast amounts of data.

Many firms such as eBay [23], Amazon [24], Walmart [3], Facebook [25], FICO [26] and so on has established a number of large-scale data warehouses to store their business data. “It estimates that the volume of business data worldwide, across all companies, doubles every 1.2 years” [27][28].

“In 2012, the Obama administration announced the Big Data Research and Development Initiative to explore how Big Data could be used to address important problems faced by the government” [16]. “This initiative is composed of 84 different Big Data programs spread across six departments” [17]. “Big Data analysis played a large role in Barack Obama's successful 2012 re-election campaign” [18]. “The Utah Data Center is a data center currently being constructed by the United States National Security Agency. When finished, the facility will be able to handle a large amount of information collected by the NSA over the Internet. The exact amount of storage space is unknown, but more recent sources claim it will be on the order of a few exabytes” [19][20][21]. “Big Data analysis was, in parts, responsible for the BJP and its allies to win a highly successful Indian General Election 2014” [22].

Organizations in any industry can benefit from Big Data [29]. “Research on the effective usage of information and communication technologies for development suggests that Big Data technology can make important contributions” [30][31]. “Advancements in Big Data analysis offer cost-effective opportunities to improve decision-making in critical development areas such as health care, employment, economic productivity, crime, security, and natural disaster and resource management” [32][33]

The McKinsey & Company report shows that Big Data may create tremendous value for the global economy through research on the five industries, i.e., the U.S. healthcare, the EU public sector administration, the U.S. retail, the global manufacturing, and the global personal location data that represent the global economy [9]. The Massachusetts Institute of Technology research indicates that companies that make decision based on data enjoy about 5% increase in productivity [29].

There is many examples of the benefits of Big Data. A notable example is Google, which found that frequency of search queries that were correlated to the time and place of the flu spreading during the spreading of flu would be different from those at ordinary times, obtained timely information that predicted and located outbreaks of the flu by aggregating massive search queries during the H1N1 crisis struck in 2009 [34]. Another example is Farecast, a technology start-up companies in the U.S, which had been purchased by Microsoft in 2008 has an airline ticket forecast system which has been incorporated into the Bing search engine that predicts the trends and increasing or decreasing range of airline ticket price. “By 2012 the system was making the correct call 75 percent of the time and saving travellers, on average, $ 50 per ticket” [7].These two examples show the scientific and societal benefits of Big Data as well as Big Data can become a source of economic value.

The most famous example of the enormous potential of Big Data may be Internet. Big Data applications in the field of Internet have e-commerce, online ad, network news, search engines, travel reservation, instant messaging, social networking, blog, microblog, online video, online music, and online game. The Internet giants such as Google, Facebook, Microsoft, Alibaba, and Tencent and so on have enormous economic potential because these firms have amassed previously unimaginable amounts of users’ data. For example, “Facebook has more than 900 million users who upload more than 250 million photos and click the “Like” button more than 2.5 billion times per day” [3].Now, these Internet companies which have already possessed massive user data are seeking new ways to utilize their customer data. And more and more organizations are collecting online user information through the construction of Big Data platform in order to share the benefit of

Big Data.

2.3 Framework and technology of Big Data

Big Data generally involves data acquisition and preparation, storage and management, analysis and mining, and interpretation. In the dig data era, with growing computational capabilities, it is feasible to use more computational power to do the same work. However, high performance network capacity has not increased at the same rate as processing and storage capabilities. The limitation in computation has simply shifted from moving data to a big supercomputer, to moving the application to many smaller computers which store the data.

Various frameworks and file systems have been developed for managing and analyzing Big Data. Typical examples are MapReduce, Hadoop and NoSQL. MapReduce framework which introduced by Google in 2004 is a programming model for producing and processing large data sets [11]. MapReduce is composed of two processes, which are Map (simple computation) and Reduce (integration) [35][36]. Hadoop is an acronym for "High-Availability Distributed Object-Oriented Platform" and is an open source implementation of MapReduce [37]. NoSQL is a technique for handling data which is difficult to handle with traditional SQL [38].

Recent studies show that the use of a multiple layer architecture is an option for dealing with Big Data" [39].A seven layers framework in which Big Data applications can be developed is proposed in [40]. This framework can be summarized in three stages, i.e., Multiple Data Sources, Data Analysis and Modelling, and Data Organization and Interpretation. Stage 1 is concerned with acquisition and filtering of data by applying correct metadata and processes. Stage 2 uses the information prepared in Stage 1 to apply analytics and predictive models to find relationships and patterns that were not initially known. Stage 3 deals with modelling the source information and mapping the data to the target model as well as interpreting the meaning of the newly discovered information.

A seven layers framework in which big data applications can be developed is proposed in [40]. This framework can be summarized in three stages, i.e., Multiple Data Sources, Data Analysis and Modelling, and Data Organization and Interpretation. Stage 1 is concerned with acquisition and filtering of data by applying correct metadata and processes. Stage 2 uses the information prepared in Stage 1 to apply analytics and predictive models to find relationships and patterns that were not initially known. Stage 3 deals with modelling the source information and mapping the data to the target model as well as interpreting the meaning of the newly discovered information.

McKinsey & Company report suggests suitable technologies to efficiently process mass data within tolerable elapsed times in Big Data include: "A/B testing [41], crowdsourcing [42], data fusion and integration [43], genetic algorithms [44], machine learning [45], natural language processing [46], signal processing [47], simulation, time series analysis [48], visualization [49][9],massively parallel-processing (MPP) databases [50], search-based applications [51], data mining [52], distributed file systems [53], distributed databases [54], cloud based infrastructure (applications, storage and computing resources) [55] and the Internet

2.4 Challenges of Big Data

The application of Big Data is leading to a set of new challenges since data sets of Big Data so large and complex that it is difficult to acquisition, storage, management and analysis. The main challenges are listed as following [56][57]:

1. Data preparation. According to the definition of strong and accurate techniques for Big Data, an important basis of Big Data analysis and management is the availability of high-quality, precise, and trustworthy data. Data preparation is paramount for increasing the value of Big data.
2. Efficient distributed storage and search. Timeliness of data collection is fundamental to provide fast analysis of Big Data. Therefore, there is an increasing need of providing efficient distributed

storage with faster memories and enhancing search algorithms.

3. Effective online data analysis. Online analysis of multidimensional data becomes a must and potential source of information for decision making. This would require adapting existing OLAP approaches to big data.
4. Effective machine learning techniques for Big Data mining. Machine learning and data mining should be adapted to Big Data to unleash the full potential of collected data.
5. Efficient handling of Big Data streams. Some specific scenarios (e.g., stock exchange) would require analysis of data in the form of streams. Fast and optimized solutions should be developed to make inference on Big Data streams.
6. Semantic lifting techniques. Semantics of collected big data represents an important aspect for future development of Big Data applications. Future approaches to Big Data analysis should be able to cope with their semantics.
7. Programming models . Various programming models of Big Data infrastructures are available. Some examples include MapReduce and Hadoop. We should consider different approaches for storing and managing data.
8. Social analytics. The ability to distinguish those data that can be trusted and comply with users' needs and preferences is important as well as different to achieve. Social analytics should then address this problem providing accurate and sound approaches to social data analysis.
9. Security and privacy. Big Data are a priceless source of information. However, it often contains sensitive information that needs to be protected from unauthorized access and release.

III. Security and Privacy

3.1 Challenges of Big Data security and privacy

Probably the most challenging and concerned problem in Big Data is security and privacy. Governmental agencies, the health care industry, biomedical researchers, and private businesses invest enormous resources into the collection, aggregation, and sharing of large amounts of personal data for the enormous benefit of Big Data. “Through recent disclosure, the National Security Administration routinely collects and analyzes massive amounts of personal data derived from heterogeneous data sources such as telecommunications, the Internet, and the user databases of large businesses, including Microsoft, Yahoo, Google, Facebook, PalTalk, YouTube, Skype, AOL, and Apple” [19]. Many facts show that Big Data will harm the user's privacy if it is not properly handled.

The security and privacy issues which should be concerned in Big Data context include: “1.The personal information of a person when combined with external large data sets leads to the inference of new facts about that person and it’s possible that these kinds of facts about the person are secretive and the person might not want the Data Owner to know or any person to know about them; 2.Information regarding the users (people) is collected and used in order to add value to the business of the organization. This is done by creating insights in their lives which they are unaware of; 3.Another important consequence arising would be Social stratification where a literate person would be taking advantages of the Big data predictive analysis and on the other hand underprivileged will be easily identified and treated worse; 4.Big Data used by law enforcement will increase the chances of certain tagged people to suffer from adverse consequences without the ability to fight back or even having knowledge that they are being discriminated” [59].

The field of privacy in big data which contains a bunch of challenges involves interaction with individuals, re-identification attacks, probable and provable results, and economic effects [60]. Interaction with individuals includes providing transparency, getting consent, revocation of consent and deletion of personal data. Re-identification attacks which have three sub-categories named correlation attacks, arbitrary identification attacks, and targeted identification attacks mean that a huge dataset available is explicitly scanned for correlations that lead to a unique fingerprint of a single individual. Probable and provable results refer the validity of the results gathered in big data. Economic effects of the big data paradigm are direct results of the exchange of datasets among business partners in advance.

3.2 Big data security techniques

Organizations used various methods of de-identification to ensure security and privacy. The most common solution to ensure security and privacy may be oral and written pledges. However, history has shown that this method is flawed. Passwords, controlled access, and two-factor authentication is low-level, but routinely used, technical solution to enforce security and privacy when sharing and aggregating data across dynamic, distributed data systems. Access permissions such as these can potentially be broken by both the intentional sharing of permissions and the continuation of permissions after they are no longer required or permitted. More advanced technological solution is cryptography. The famous encryption schemes have AES and RSA. Recent revelations show that the National Security Administration (NSA) may have already found ways to break or circumvent existing Internet encryption schemes [61]. Virtual barriers such as firewalls, secure sockets layer and transport layer security are designed to restrict access to data. Each of these technologies can be broken, however, and thus need to be constantly monitored, with fixes applied as needed. Tracking, monitoring or auditing software is developed to provide a history of data flow and network access by an individual user in order to ensure compliance with security-related. The limitation of this technology is that it is difficult and costly to implement on a large scale or with distributed data systems and users because it requires dedicated staff to read and interpret the findings, and the software can be exploited to monitor individual behavior rather than protecting data.All in all, the traditional de-identification techniques are not applicable in the era of Big Data since the de-identification technique widespread uses. The tasks of ensuring Big Data security and privacy become more difficult as information is increased. "Computer scientists have repeatedly shown that even anonymized data can often be re-identified and attributed to specific individuals" [62].

A novel technological named the integrated Rule-Oriented Data (iRODS) is proposed to be the solution to ensure security and privacy in big data [60]. iRODS was architected and designed to address these challenges across a broad spectrum of communities, with differing institutional goals and security and privacy concerns, by providing each adopter community the ability to develop and deploy solutions for data management and sharing that are specific to organizational needs [63][64] Key technological features of iRODS include: federated data grids or "intelligent clouds", a distributed rules engine, an "iCAT" metadata catalog, a storage access layer that allows common access, a rich combination of graphical user interface and command-line–based clients and APIs for interaction with an iRODS data grid. iRODS is used in a number of data management applications and has been adopted by numerous institutions around the world. Many publications describe the myriad ways in which iRODS technology has been adapted and applied to solve of variety of challenges in policy based, large-scale data management [63][64][65][66].The iRODS technology provides improvements in common approaches to securing data and ensuring privacy, including: comprehensive set of security controls, improved control of data access and use through metadata, storage virtualization and data security lifecycle, and persistent identifiers.

The big data security techniques data also involve released anonymity protection, social networking anonymity protection and data provenance and so on.

For structured data in big data, the data released anonymity protection which is the basic means and key technologies to achieve protection of privacy are still in the stage of continuous development and improvement. The early [67][68]and optimal [69][70][71] k-anonymity protecting scheme focus on static and one-time data released situation. In reality data released often faced continuous and repeatedly situation. In contrast, the data released anonymity protection is more sophisticated in big data scene.

Social networking is an important source of data in big data, while these data contains a large number of user privacy data. Social networks anonymity protection need to implement anonymous for user ID. The important issue of social networking anonymous protection is the attacker may be inferred anonymous user based on the connection between the users through other publicly available information. For example, various link prediction algorithms have been proposed in [72][73][74][75]. Study has shown that the gather characteristics of social networking for accurate relationship forecasting methods have an important influence [76]. Therefore, the future anonymity protection technology should be able to effectively resist such speculation attacks.

Data provenance technology has been widely studied in the database field before the emergence of the concept of big data [77]. In the national cyberspace security report in 2009, data provenance is listed as one of security key technologies which ensure the future nation's critical infrastructure [78]. The balance between data provenance and privacy protection and the security of data provenance technology are challenges to be faced when data provenance technology is employed in big data security and privacy protection.

We should be mindful of sizable benefits of Big Data and the cost of implementing security and privacy when considering the risks of Big Data security and privacy. However, the benefits of Big Data do not always belong to the users whose personal data are collected. A legal frame where the benefits of data for organizations are shared with individuals are discussed in [79].

Notably, many of security and privacy challenges do not stem from technical issues, but merely are based on legislation and organizational matters. “Both privacy advocates and security experts agree that the laws governing electronic eavesdropping have not kept pace with technology” [80]. In a survey of technology and legal Issues of Big Data and NSA surveillance describes legal issues for NSA’s activities with Big Data, legal issues surfaced by disclosure of NSA surveillance activities, cooperation of private enterprises and NSA according to common goals, technical and legal analysis of details for NSA’s activities based on publicized materials, and opinions regarding NSA surveillance activity [81].

IV. Conclusion

Security and privacy are among the most important requirements in Big Data. We have seen that the best solution of implementing Big Data security and privacy is the law rather than the security technology but the laws can’t keep pace with the development of technology and is different between countries. Thus, security technology and other methods are always necessary. Some possible methods and techniques to ensure security and privacy in Big Data have been discussed above. Moreover, we noted that correlation of massive data is the key which is the basis of use of Big Data as well as the reason of Big Data security and privacy issues. We recommend that the study of “no correlation” may be realization of security and privacy in Big Data.

Acknowledgements

We gratefully acknowledge anonymous reviewers who read drafts and made many helpful suggestions. This work is supported by the National Natural Science Foundation of China under Grant No. 61202079; the China Postdoctoral Science Foundation under Grant No. 2013M530526; and the Fundamental Research Funds for the Central Universities under Grant No. FRF-TP-14-046A2.

References

[1] Gantz J, Reinsel D. Extracting value from chaos [J]. IDC iview, 2011: 1-12.

[2] Lohr S. The age of big data [J]. New York Times, 2012, 11.

[3] McCune J C. Data, data everywhere [J]. Management Review, 1998, 87(10): 10-12.

[4] Gantz J, Reinsel D. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the Far East [J]. IDC iView: IDC Analyze the Future, 2012.

[5] Weiss R, Zgorski L. Obama Administration Unveils “Big Data” Initiative: Announces $200 Million in New R&D Investments [J]. Office of Science and Technology Policy, Washington, DC, 2012. Data P. The Emergence of a New Asset Class[C]//World Economic Forum Report. 2011.

[6] Anderson C. The end of theory: the data deluge makes the scientific method obsolete. Wired Magazine 16.07[J]. 2008.

[7] Mayer-Schönberger V, Cukier K. Big data: A revolution that will transform how we live, work, and think [M]. Houghton Mifflin Harcourt, 2013.

[8] Ardagna C A, Damiani E. Business Intelligence meets Big Data: An Overview on Security and Privacy [J].

[9] Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity [J]. 2011.

[10] Laney D. 3-D Data Management: Controlling Data Volume [J]. Velocity and Variety, META Group Original Research Note, 2001.

[11] Beyer M. Gartner says solving big data challenge involves more than just managing volumes of data. Gartner [J]. 2011.

[12] Beyer M A, Laney D. The importance of 'big data': a definition [J]. Stamford, CT: Gartner, 2012.

[13] Lefevre C. LHC: the guide (English version) [R]. 2009.

[14] Brumfiel G. Down the petabyte highway[J]. Nature, 2011, 469(20): 282-283.

[15] Mangelsdorf J. Supercomputing the climate: Nasa’s big data mission[J]. Accessed online, 2013: 11-27.

[16] Kalil T. Big data is a big deal[J]. The White House, 2012.

[17] Sheet F. Big Data Across the Federal Government[J]. 2012 03-29)[2013-03-06]. http://www. whitehouse, gov/sites/default/files/microsites/ostp/big_ data fact sheet final. pdf, 2012.

[18] Lampitt A. ’The real story of how Big Data analytics helped Obama win’[J]. Info World, 2013, 14.

[19] Bamford J. The NSA is building the country’s biggest spy center (Watch What You Say)[J]. Wired, March, 2012, 15.

[20] Groundbreaking Ceremony Held for $1.2 Billion Utah Data Center. National Security Agency Central Security Service. Retrieved 2013

[21] Hill , Kashmir. TBlueprints Of NSA's Ridiculously Expensive Data Center In Utah Suggest It Holds Less Info Than Thought. Forbes. Retrieved 2013-10-31.

[22] News: Live Mint. Are Indian companies making enough sense of Big Data?. Live Mint - http://www.livemint.com/. 2014-06-23. Retrieved 2014-11-22.

[23] Tay L. Inside eBay’s 90PB data warehouse[J]. 2010.

[24] Layton J. Amazon Technology[J]. 2013--03--05)[-2013--06--06]. Money. howstuffworks, com.

[25] Johnson R. Scaling facebook to 500 million users and beyond[J]. Retrieved May, 2010, 4: 2014.

[26] FICO® Falcon® Fraud Manager. http://www.fico.com/en/products/fico-falcon-fraud-manager/

[27] eBay Study: How to Build Trust and Improve the Shopping Experience. Knowwpcarey.com. 2012-05-08

[28] Leading Priorities for Big Data for Business and IT. eMarketer. October 2013

[29] Brynjolfsson E, Hitt L, Kim H. Strength in Numbers: How does data-driven decision-making affect firm performance? [J]. 2011.

[30] Pulse U N G. Big Data for Development: Opportunities & Challenges [J]. White paper, Global pulse, New York, USA, May, 2012.

[31] Data B. Big Impact: New Possibilities for International Development [J]. 2013-04-07].

http://www3. weforum, org/docs/WEF-TC-MFS-BigData BigImpact_Briefing. _2012. pdf.

[32] Hilbert M. Big data for development: From information-to knowledge societies [J]. Available at SSRN 2205145, 2013.

[33] Elena Kvochko, Four Ways To talk About Big Data (Information Communication Technologies for Development Series). worldbank. org.

[34] Carneiro H A, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks [J]. Clinical infectious diseases, 2009, 49(10): 1557-1564.

[35] Yang H, Dasdan A, Hsiao R L, et al. Map-reduce-merge: simplified relational data processing on large clusters[C]//Proceedings of the 2007 ACM SIGMOD international conference on Management of data. ACM, 2007: 1029-1040.

[36] Cohen J. Graph twiddling in a MapReduce world[J]. Computing in Science & Engineering, 2009, 11(4): 29-41.

[37] Center Intel IT. Planning Guide: Getting Started with Hadoop [J]. Steps IT Managers Can Take to Move Forward with Big Data Analytics, 2012.

[38] Bonnet L, Laurent A, Sala M, et al. Reduce, you say: What nosql can do for data aggregation and bi in large repositories[C]// Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on. IEEE, 2011: 483-488.

[39] Boja C, Pocovnicu A, Batagan L. Distributed Parallel Architecture for Big Data[J]. Informatica Economica, 2012, 16(2): 116-127.

[40] Tekiner F, Keane J A. Big Data Framework [C]//Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on. IEEE, 2013: 1494-1499.

[41] Christian B. The A/B test: Inside the technology that's changing the rules of business[J]. Wired Magazine, 2012.

[42] Howe J. The rise of crowdsourcing[J]. Wired magazine, 2006, 14(6): 1-4.

[43] Klein L A. Sensor and data fusion: a tool for information assessment and decision making[M]. Bellingham^ eWA WA: Spie Press, 2004.

[44] Eiben A E, Raue P E, Ruttkay Z. Genetic algorithms with multiparent recombination[M]//Parallel Problem Solving from Nature—PPSN III. Springer Berlin Heidelberg, 1994: 78-87.

[45] Witten I H, Frank E. Data Mining: Practical machine learning tools and techniques[M]. Morgan Kaufmann, 2005.

[46] Manning C D. Foundations of statistical natural language processing[M]. MIT press, 1999.

[47] Lyons R G. Understanding digital signal processing[M]. Pearson Education, 2010.

[48] Hamilton J D. Time series analysis[M]. Princeton: Princeton university press, 1994.

[49] Van Dam A, Feiner S K. Computer graphics: principles and practice[M]. Pearson Education, 2014.

[50] Novick I D, Subramanya A B, Devarayasamudram Nagendran S, et al. PARTITION LEVEL BACKUP AND RESTORE OF A MASSIVELY PARALLEL PROCESSING DATABASE: U.S. Patent 20,140,310,245[P]. 2014-10-16.

[51] Harman M, Jones B F. Search-based software engineering[J]. Information and Software Technology, 2001, 43(14): 833-839.

[52] Fayyad U M, Piatetsky-Shapiro G, Smyth P, et al. Advances in knowledge discovery and data mining[J]. 1996.

[53] Satyanarayanan M. Distributed file systems[J]. Distributed Systems. Addison-Wesley and ACM Press,, 1993, 821: 145-154.

[54] Rothnie Jr J B, Bernstein P A, Fox S, et al. Introduction to a system for distributed databases (SDD-1)[J]. ACM Transactions on Database Systems (TODS), 1980, 5(1): 1-17.

[55] Kloch C, Petersen E B, Madsen O B. Cloud based infrastructure, the new business possibilities and barriers[J]. Wireless Personal Communications, 2011, 58(1): 17-30.

[56] Ardagna C A, Damiani E. Business Intelligence meets Big Data: An Overview on Security and Privacy[J].

[57] Labrinidis A, Jagadish H V. Challenges and opportunities with big data [J]. Proceedings of the VLDB Endowment, 2012, 5(12): 2032-2033.

[58] Schmitt C. Security and Privacy in the Era of Big Data[J].

[59] Katal A, Wazid M, Goudar R H. Big data: Issues, challenges, tools and Good practices[C]//Contemporary Computing (IC3), 2013 Sixth International Conference on. IEEE, 2013: 404-409.

[60] Jensen M. Challenges of privacy protection in big data analytics[C]//Big Data (BigData Congress), 2013 IEEE International Congress on. IEEE, 2013: 235-238.

[61] Perlroth N, Larson J, Shane S. NSA able to foil basic safeguards of privacy on web[J]. The New York Times, 2013, 6.

[62] Ohm P. Broken promises of privacy: Responding to the surprising failure of anonymization [J]. UCLA L. Rev., 2009, 57: 1701.

[63] Rajasekar A, Moore R, Hou C, et al. iRODS Primer: integrated rule-oriented data system[J]. Synthesis Lectures on Information Concepts, Retrieval, and Services, 2010, 2(1): 1-143.

[64] Rajasekar A, Moore R, Wan M, et al. Applying rules as policies for large-scale data sharing[C]//Intelligent Systems, Modelling and Simulation (ISMS), 2010 International Conference on. IEEE, 2010: 322-327.

[65] Barg I, Scott D, Timmermann E. NOAO E2E integrated data cache initiative using iRODS[C]//Astronomical Data Analysis Software and Systems XX. ASP Conference Proceedings. 2011, 442: 497-500.

[66] Schnase J L, Webster W P, Parnell L A, et al. The NASA Center for Climate Simulation Data Management System[C]//Mass Storage Systems and Technologies (MSST), 2011 IEEE 27th Symposium on. IEEE, 2011: 1-6.

[67] Sweeney L. k-anonymity: A model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(05): 557-570.

[68] Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(05): 571-588.

[69] Bayardo R J, Agrawal R. Data privacy through optimal k-anonymization[C]//Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on. IEEE, 2005: 217-228.

[70] LeFevre K, DeWitt D J, Ramakrishnan R. Incognito: Efficient full-domain k-anonymity[C]//Proceedings of the 2005 ACM SIGMOD international conference on Management of data. ACM, 2005: 49-60.

[71] LeFevre K, DeWitt D J, Ramakrishnan R. Mondrian multidimensional k-anonymity[C]//Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on. IEEE, 2006: 25-25.

[72] Lü L, Zhou T. Link prediction in weighted networks: The role of weak ties[J]. EPL (Europhysics Letters), 2010, 89(1): 18001.

[73] Clauset A, Moore C, Newman M E J. Hierarchical structure and the prediction of missing links in networks[J]. Nature, 2008, 453(7191): 98-101.

[74] Yin D, Hong L, Xiong X, et al. Link formation analysis in microblogs[C]//Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011: 1235-1236.

[75] Lichtenwalter R N, Lussier J T, Chawla N V. New perspectives and methods in link prediction[C]//Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010: 243-252.

[76] Lü L, Zhou T. Link prediction in complex networks: A survey[J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150-1170.

[77] Buneman P, Khanna S, Wang-Chiew T. Why and where: A characterization of data provenance[M]//Database Theory—ICDT 2001. Springer Berlin Heidelberg, 2001: 316-330.

[78] Wynbourne M N, Austin M F, Palmer C C. National Cyber Security Research and Development Challenges Related to Economics, Physical Infrastructure and Human Behavior: An Industry, Academic and Government Perspective[M]. Institute for Information Infrastructure Protection, 2009.

[79] Tene O, Polonetsky J. Privacy in the age of big data: A time for big decisions[J]. Stanford Law Review Online, 2012, 64: 63.

[80] Tene O, Polonetsky J. Big data for all: Privacy and user control in the age of analytics[J]. 2013.

[81] Park C, Wang T. Big Data and NSA Surveillance--Survey of Technology and Legal Issues[C]//Multimedia (ISM), 2013 IEEE International Symposium on. IEEE, 2013: 516-517.

Biographies

Matturdi Bardi graduated from Mathematics and Applied Mathematics of Hotan Teachers College in July 1982 and still stay teaching, in 1993 graduated from the Mathematics and Applied Mathematics (undergraduate) of Xinjiang Institute of Education. Meanwhile engaged in advanced studies in East China Normal University, Xinjiang Normal University and Capital Normal Universityfor further study. 1995 to 2013, served as the Director of math department in Hotan Teachers College .Currently serve as Secretary, vice president, associate professor of School of Mathematics and Information in Hotan Teachers College .Email: matturdibardi@163.com

ZHOU Xianwei received his B.S. degree in Department of Mathematics from Southwest Normal University in 1986 and his M.S. degree from Zhengzhou University in 1992, and in 1999, he obtained Ph.D. degree in Department of Transportation Engineering from Southwest Jiaotong University, China. He was engaged in postdoctor study at Beijing Jiaotong University, China, from 1999 to 2000. Now, as a professor in Department of Communication Engineering, School of Computer and Communication Engineering, University of Science and Technology Beijing, his research interests include the security of communication networks, next generation networks, scheduling and game theory.

LI Shuai, received his M.S. degrees from School of Mathematical Sciences, Soochow University , China in 2013. He is currently PH.D in the Department of Computer and Communication Engineering, University of Science and Technology Beijing, China. His research interests is communications security. Email: mrzry@163.com

LIN Fuhong received his M.S. degree and Ph.D degree from Beijing Jiaotong University, Beijing, China, in 2006 and 2010, respectively, both in Electronics Engineering. Now he is a lecturer in department of Computer and Communication Engineering, University of Science and Technology Beijing, P. R. China. His research interests include wisdom network, social network and multimedia network. *The corresponding author. Email: FHLin_ustb@yeah.net