The Development and Application of Big Data

Mimi Yu [1, a, *]

[1]School of Management, Shanghai University, Shanghai 200444, China.

[a]Corresponding author e-mail: doublemi_yu@163.com

Abstract

We are constantly being told that we live in the Information Era – the Age of BIG data. The speed of development in Big Data and associated phenomena, such as social media, has surpassed the capacity of the average consumer to understand his or her actions and their knock-on effects. This paper first introduces the definition, features and challenges of big data. Then this paper examines the several representative applications of big data, including enterprise management, Internet of Things, online social networks, medial applications, collective intelligence, and smart grid. Finally, this paper introduces the conclusions and outlook of the big data.

Keywords

Big data; development; challenges; application.

1. Introduction

Big data is all the rage. Its proponents tout the use of sophisticated analytics to mine large data sets for insight as the solution to many of our society's problems. These big data evangelists insist that data-driven decision making can now give us better predictions in areas ranging from college admissions to dating to hiring. Big Data is everywhere. In recent years, there is an increasing emphasis on big data, business analytics and 'smart' living and work environments. We don't deny that big data holds substantial potential for the future, and that large dataset analysis has important uses today.

Big Data is perceived as "the new driver of competitive advantage" [1]. Big Data applications have high Volume, high Variety and high Velocity as findings are expected to be delivered very quickly [2]. In reality, Big Data is largely driven by the need to analyze massive volumes of data to gain competitive advantage and to use previously intractable processes to find information/ relationships [3]. This also provides more opportunities to bring in applications and fields to enrich data further to provide even better analysis.

The big data has been widely used in Public Management Administration, medical services, retail, manufacturing, and location services involving individuals, and has produced great social value and industrial space. The large commercial value, scientific research value, social management and public service value and the value of supporting scientific decision-making in big data are being recognized and exploited. Therefore, the development and application of big data, not only has a prominent scientific frontier and significant strategic significance, but also has great practical value and distinctive characteristics of the times.

2. The Development of Big Data

2.1. Definition and Features of Big Data

Big data is an abstract concept. In general, big data shall mean the datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time. Because of different concerns, scientific and technological enterprises, research scholars, data analysts, and technical practitioners have different definitions of big

data. The following definitions may help us have a better understanding on the profound social, economic, and technological connotations of big data.

Big data shall mean such datasets which could not be acquired, stored, and managed by classic database software. This definition includes two connotations: First, datasets' volumes that conform to the standard of big data are changing, and may grow over time or with technological advances; Second, datasets' volumes that conform to the standard of big data in different applications differ from each other. At present, big data generally ranges from several TB to several PB[1]. From the definition by McKinsey & Company, it can be seen that the volume of a dataset is not the only criterion for big data. The increasingly growing data scale and its management that could not be handled by traditional database technologies are the next two key features.

In 2011, an IDC report 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." [4].With this definition, characteristics of big data may be summarized as four Vs, i.e., Volume (great volume), Variety (various modalities), Velocity (rapid generation), and Value (huge value but very low density), as shown in Figure 1. Such 4Vs definition was widely recognized since it highlights the meaning and necessity of big data, i.e., exploring the huge hidden values. This definition indicates the most critical problem in big data, which is how to discover values from datasets with an enormous scale, various types, and rapid generation. As Jay Parikh, Deputy Chief Engineer of Facebook, said, "You could only own a bunch of data other than big data if you do not utilize the collected data."[5].

In addition to developing a proper definition, the big data research should also focus on how to extract its value, how to use data, and how to transform "a bunch of data" into "big data."

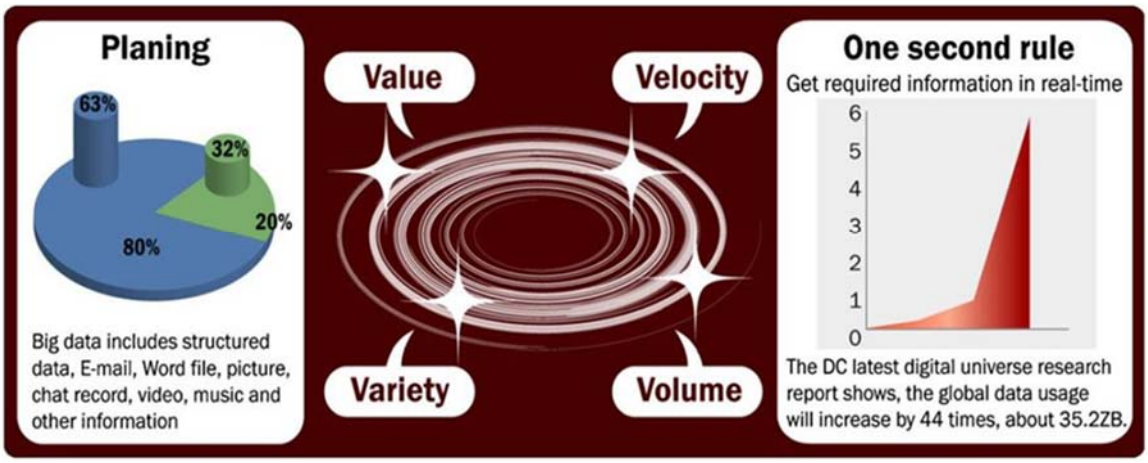

Figure 1. The 4Vs feature of big data

2.2. Definition a Development of Big Data

In the late 1970s, the concept of "database machine" emerged, which is a technology specially used for storing and analyzing data. In the 1980s, people proposed "share nothing," a parallel database system, to meet the demand of the increasing data volume [6].

In the 1980s, people proposed "share nothing," a parallel database system, to meet the demand of the increasing data volume [6].

On June 2, 1986, a milestone event occurred when Teradata delivered the first parallel database system with the storage capacity of 1TB to Kmart to help the large-scale retail company in North America to expand its data warehouse [7].

In the late 1990s, the advantages of parallel database was widely recognized in the database field. In January 2007, Jim Gray, a pioneer of database software, called such transformation "The Fourth Paradigm" [8]. He also thought the only way to cope with such paradigm was to develop a new generation of computing tools to manage, visualize, and analyze massive data. In June 2011, another milestone event occurred; EMC/IDC published a research report titled Extracting Values from Chaos [4], which introduced the concept and potential of big data for the first time. This research report triggered the great interest in both industry and academia on big data.

Over the past few years, nearly all major companies, including EMC, Oracle, IBM, Microsoft, Google, Amazon, and Facebook, etc. have started their big data projects.

In 2008, Nature published a big data special issue. In 2011, Science also launched a special issue on the key technologies of "data processing" in big data. In 2012, European Research Consortium for Informatics and Mathematics (ERCIM) News published a special issue on big data.

In the beginning of 2012, a report titled Big Data, Big Impact presented at the Davos Forum in Switzerland, announced that big data has become a new kind of economic assets, just like currency or gold. Gartner, an international research agency, issued Hype Cycles from 2012 to 2013, which classified big data computing, social analysis, and stored data analysis into 48 emerging technologies that deserve most attention.

Many national governments such as the U.S. also paid great attention to big data. In March 2012, the Obama Administration announced a USD 200 million investment to launch the "Big Data Research and Development Plan," which was a second major scientific and technological development initiative after the "Information Highway" initiative in 1993. In July 2012, the "Vigorous ICT Japan" project issued by Japan's Ministry of Internal Affairs and Communications indicated that the big data development should be a national strategy and application technologies should be the focus. In July 2012, the United Nations issued Big Data for Development report, which summarized how governments utilized big data to better serve and protect their people.

2.3. Challenges of Big Data

The sharply increasing data deluge in the big data era brings about huge challenges on data acquisition, storage, management and analysis. Traditional data management and analysis systems are based on the relational database management system (RDBMS). However, such RDBMSs only apply to structured data, other than semi-structured or unstructured data. In addition, RDBMSs are increasingly utilizing more and more expensive hardware. It is apparently that the traditional RDBMSs could not handle the huge volume and heterogeneity of big data. The research community has proposed some solutions from different perspectives. For example, cloud computing is utilized to meet the requirements on infrastructure for big data, e.g., cost efficiency, elasticity, and smooth upgrading/downgrading. For solutions of permanent storage and management of large-scale disordered datasets, distributed file systems [9] and NoSQL [10] databases are good choices. Such programming frameworks have achieved great success in processing clustered tasks, especially for webpage ranking. Various big data applications can be developed based on these innovative technologies or platforms. Moreover, it is non-trivial to deploy the big data analysis systems. Some literatures [13][13] discuss obstacles in the development of big data applications. The key challenges are listed as follows:

1. Data representation: many datasets have certain levels of heterogeneity in type, structure, semantics, organization, granularity, and accessibility. Data representation aims to make data more meaningful for computer analysis and user interpretation. Nevertheless, an improper data representation will reduce the value of the original data and may even obstruct effective

data analysis. Efficient data representation shall reflect data structure, class, and type, as well as integrated technologies, so as to enable efficient operations on different datasets [14].

2. Redundancy reduction and data compression: generally, there is a high level of redundancy in datasets. Redundancy reduction and data compression is effective to reduce the indirect cost of the entire system on the premise that the potential values of the data are not affected. For example, most data generated by sensor networks are highly redundant, which may be filtered and compressed at orders of magnitude.

3. Data life cycle management: compared with the relatively slow advances of storage systems, pervasive sensing and computing are generating data at unprecedented rates and scales. We are confronted with a lot of pressing challenges, one of which is that the current storage system could not support such massive data. Generally speaking, values hidden in big data depend on data freshness. Therefore, a data importance principle related to the analytical value should be developed to decide which data shall be stored and which data shall be discarded.

4. Analytical mechanism: the analytical system of big data shall process masses of heterogeneous data within a limited time. However, traditional RDBMSs are strictly designed with a lack of scalability and expandability, which could not meet the performance requirements. Non-relational databases have shown their unique advantages in the processing of unstructured data and started to become mainstream in big data analysis.

5. Data confidentiality: most big data service providers or owners at present could not effectively maintain and analyze such huge datasets because of their limited capacity. They must rely on professionals or tools to analyze such data, which increase the potential safety risks. For example, the transactional dataset generally includes a set of complete operating data to drive key business processes. Such data contains details of the lowest granularity and some sensitive information such as credit card numbers. Therefore, analysis of big data may be delivered to a third party for processing only when proper preventive measures are taken to protect such sensitive data, to ensure its safety.

6. Energy management: the energy consumption of mainframe computing systems has drawn much attention from both economy and environment perspectives. With the increase of data volume and analytical demands, the processing, storage, and transmission of big data will inevitably consume more and more electric energy. Therefore, system-level power consumption control and management mechanism shall be established for big data while the expandability and accessibility are ensured.

7. Expendability and scalability: the analytical system of big data must support present and future datasets. The analytical algorithm must be able to process increasingly expanding and more complex datasets.

8. Cooperation: analysis of big data is an interdisciplinary research, which requires experts in different fields cooperate to harvest the potential of big data. A comprehensive big data network architecture must be established to help scientists and engineers in various fields access different kinds of data and fully utilize their expertise, so as to cooperate to complete the analytical objectives.

3. Big Data Applications

Data analysis involves a wide range of applications, which frequently change and are extremely complex [15].

3.1. Application of Big Data in Enterprises

At present, big data mainly comes from and is mainly used in enterprises, while BI and OLAP can be regarded as the predecessors of big data application. The application of big data in enterprises can enhance their production efficiency and competitiveness in many aspects [16].

In particular, on marketing, with correlation analysis of big data, enterprises can more accurately predict the consumer behavior and find new business modes. On sales planning, after comparison of massive data, enterprises can optimize their commodity prices. On operation, enterprises can improve their operation efficiency and satisfaction, optimize the labor force, accurately forecast personnel allocation requirements, avoid excess production capacity, and reduce labor cost. On supply chain, using big data, enterprises may conduct inventory optimization, logistic optimization, and supplier coordination, etc., to mitigate the gap between supply and demand, control budgets, and improve services. In finance, the application of big data in enterprises has been rapidly developed. For example, China Merchants Bank (CMB) utilizes data analysis to recognize that such activities as "Multi-times score accumulation" and "score exchange in shops" are effective for attracting quality customers. By analyzing customers' transaction records, potential small business customers can be efficiently identified. By utilizing remote banking and the cloud referral platform to implement cross-selling, considerable performance gains were achieved [17].

Obviously, the most classic application is in e-commerce. Tens of thousands of transactions are conducted in Taobao and the corresponding transaction time, commodity prices, and purchase quantities are recorded every day, and more important, along with age, gender, address, and even hobbies and interests of buyers and sellers [15]. Data Cube of Taobao is a big data application on the Taobao platform, through which, merchants can be wared of the macroscopic industrial status of the Taobao platform, market conditions of their brands, and consumers' behaviors, etc., and accordingly make production and inventory decisions. Meanwhile, more consumers can purchase their favorite commodities with more preferable prices. The credit loan of Alibaba automatically analyzes and judges weather to lend loans to enterprises through the acquired enterprise transaction data by virtue of big data technology, while manual intervention does not occur in the entire process. It is disclosed that, so far, Alibaba has lent more than RMB 30 billion Yuan with only about 0.3 % bad loans, which is greatly lower than those of other commercial banks.

3.2. Application of IoT based Big Data

IoT is not only an important source of big data, but also one of the main markets of big data applications. Because of the high variety of objects, the applications of IoT also evolve endlessly.

Logistic enterprises may have profoundly experienced with the application of IoT big data. For example, trucks of UPS are equipped with sensors, wireless adapters, and GPS, so the Headquarter can track truck positions and prevent engine failures. Meanwhile, this system also helps UPS to supervise and manage its employees and optimize delivery routes. The optimal delivery routes specified for UPS trucks are derived from their past driving experience.

Smart city is a hot research area based on the application of IoT data. For example, the smart city project cooperation between the Miami-Dade County in Florida and IBM closely connects 35 types of key county government departments and Miami city and helps government leaders obtain better information support in decision making for managing water resources, reducing traffic jam, and improving public safety [15]. The application of smart city brings about benefits in many aspects for Dade County. Smart city advocates argue enables real-time analysis of city life, new modes of urban governance, and provides the raw material for envisioning and enacting more efficient, sustainable, competitive, productive, open and transparent cities.

3.3. Application of Online Social Network-Oriented Big Data

Online SNS is a social structure constituted by social individuals and connections among individuals based on an information network. Big data of online SNS mainly comes from instant messages, online social, micro blog, and shared space, etc, which represents various user activities. The analysis of big data from online SNS uses computational analytical method provided for understanding relations in the human society by virtue of theories and methods,

which involves mathematics, informatics, sociology, and management science, etc., from three dimensions including network structure, group interaction, and information spreading. The application includes network public opinion analysis, network intelligence collection and analysis, socialized marketing, government decision-making support, and online education, etc. Figure 2 illustrates the technical framework of the application of big data of online SNS [19].

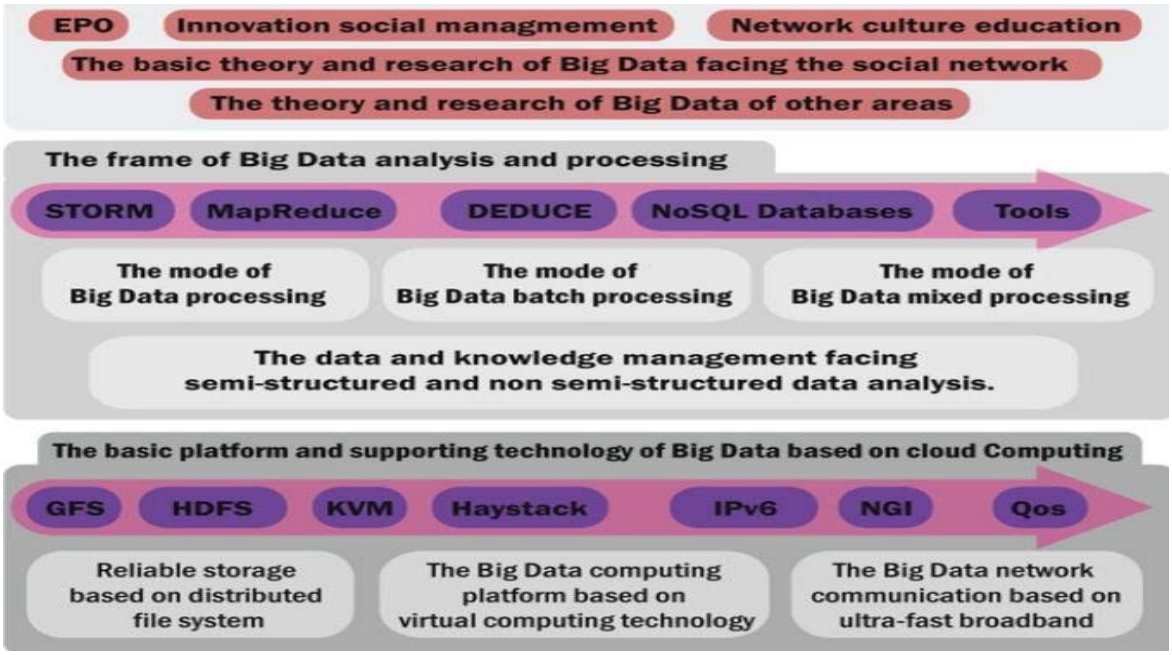

Figure 2. Enabling technologies for online social network-oriented big data

Classic applications of big data from online SNS are introduced in the following, which mainly mine and analyze content information and structural information to acquire values [18].

1. Content-based Applications: Language and text are the two most important forms of presentation in SNS. Through the analysis of language and text, user preference, emotion, interest, and demand, etc. may be revealed.

2. Structure-based Applications: In SNS, users are represented as nodes while social relation, interest, and hobbies, etc. aggregate relations among users into a clustered structure. Such structure with close relations among internal individuals but loose external relations is also called a community. The community-based analysis is of vital importance to improve information propagation and for interpersonal relation analysis.

Generally speaking, the application of big data from online SNS may help to better understand user's behavior and master the laws of social and economic activities from the following three aspects:

1. Early Warning: to rapidly cope with crisis if any by detecting abnormalities in the usage of electronic devices and services.

2. Real-time Monitoring: to provide accurate information for the formulation of policies and plans by monitoring the current behavior, emotion, and preference of users.

3. Real-time Feedback: acquire groups' feedbacks against some social activities based on real-time monitoring.

3.4. Applications of Healthcare and Medical Big Data

Healthcare and medical data are continuously and rapidly growing complex data, containing abundant and diverse information values. Big data has unlimited potential for effectively

storing, processing, querying, and analyzing medical data. The application of medical big data will profoundly influence the health care business.

For example, Aetna Life Insurance Company selected 102 patients from a pool of a thousand patients to complete an experiment in order to help predict the recovery of patients with metabolic syndrome. In an independent experiment, it scanned 600,000 laboratory test results and 180,000 claims through a series of detection test results of metabolic syndrome of patients in three consecutive years. In addition, it summarized the final result into an extreme personalized treatment plan to assess the dangerous factors and main treatment plans of patients. Then, doctors may reduce morbidity by 50 % in the next 10 years by prescribing statins and helping patients to lose weight by five pounds, or suggesting patients to reduce the total triglyceride in their bodies if the sugar content in their bodies is over 20.

3.5. Collective Intelligence

With the rapid development of wireless communication and sensor technologies, mobile phones and tablet have increasingly stronger computing and sensing capacities. As a result, crowd sensing is becoming a key issue of mobile computing. In crowd sensing, a large number of general users utilize mobile devices as basic sensing units to conduct coordination with mobile networks for distribution of sensed tasks and collection and utilization of sensed data. It can help us complete large-scale and complex social sensing tasks. In crowd sensing, participants who complete complex sensing tasks do not need to have professional skills. Crowd sensing in the form of Crowdsourcing has been successfully applied to geotagged photograph, positioning and navigation, urban road traffic sensing, market forecast, opinion mining, and other labor-intensive applications.

Crowdsourcing, a new approach for problem solving, takes a large number of general users as the foundation and distributes tasks in a free and voluntary manner. As a matter of fact, Crowdsourcing has been applied by many companies before the emergence of big data. For example, P & G, BMW, and Audi improved their R&D and design capacities by virtue of Crowdsourcing. The main idea of Crowdsourcing is to distribute tasks to general users and to complete tasks that individual users could not or do not want to accomplish. With no need for intentionally deploying sensing modules and employing professionals, Crowdsourcing can broaden the scope of a sensing system to reach the city scale and even larger scales.

In the big data era, Spatial Crowdsourcing becomes a hot topic. The operation framework of Spatial Crowdsourcing is shown as follows. A user may request the service and resources related to a specified location. Then the mobile users who are willing to participate in the task will move to the specified location to acquire related data (such as video, audio, or pictures). Finally, the acquired data will be send to the service requester. With the rapid growth of mobile devices and the increasingly powerful functions provided by mobile devices, it can be forecasted that Spatial Crowdsourcing will be more prevailing than traditional Crowdsourcing, e.g., Amazon Turk and Crowdflower.

3.6. Smart Grid

Smart Grid is the next generation power grid constituted by traditional energy networks integrated with computation, communications and control for optimized generation, supply, and consumption of electric energy. Smart Grid related big data are generated from various sources [20], such as power utilization habits of users; phasor measurement data, which are measured by phasor measurement unit (PMU) deployed national-wide; energy consumption data measured by the smart meters in the Advanced Metering Infrastructure (AMI); energy market pricing and bidding data; management, control and maintenance data for devices and equipment in the power generation, transmission and distribution networks (such as Circuit Breaker Monitors and transformers). Smart Grid brings about the following challenges on exploiting big data.

1. Grid planning: By analyzing data in the Smart Grid, the regions can be identified that have excessive high electrical load or high power outage frequencies. Even the transmission lines with high failure probability can be identified. Such analytical results may contribute to grid upgrading, transformation, and maintenance, etc. Preferential transformation on the power grid facilities in blocks with high power outage frequencies and serious overloads may be conducted, as displayed in the map.

2. Interaction between power generation and power consumption: An ideal power grid shall balance power generation and consumption. However, the traditional power grid is constructed based on one-directional approach of transmission-transformation-distribution consumption, which does not allow adjust the generation capacity according to the demand of power consumption, thus leading to electric energy redundancy and waste. Therefore, smart electric meters are developed to improve power supply efficiency. TXU Energy has several successful deployment of smart electric meters, which can help supplier read power utilization data in every 15min other than every month in the past. Labor cost for meter reading is greatly reduced, because power utilization data (a source of big data) are frequently and rapidly acquired and analyzed, power supply companies can adjust the electricity price according to peak and low periods of power consumption. TXU Energy utilized such price level to stabilize the peak and low fluctuations of power consumption. As a matter of fact, the application of big data in the smart grid can help the realization of time-sharing dynamic pricing, which is a win-win situation for both energy suppliers and users.

3. The access of intermittent renewable energy: At present, many new energy resources, such as wind and solar, can be connected to power grids. However, since the power generation capacities of new energy resources are closely related to climate conditions that feature randomness and intermittency, it is challenging to connect them to power grids. If the big data of power grids is effectively analyzed, such intermittent renewable new energy sources can be efficiently managed: the electricity generated by the new energy resources can be allocated to regions with electricity shortage. Such energy resources can complement the traditional hydropower and thermal power generations.

4. Conclusion and Outlook

This paper reviews the development and applications of big data. Firstly, we introduce the general definition and features of big data. Then the paper focuses on the challenges of the big data. The analysis of big data is confronted with many challenges, but the current research is still in early stage. Finally this paper reviews the several representative applications of big data, including enterprise management, IoT, social networks, medical applications, collective intelligence, and smart grid. These discussions aim to provide a comprehensive overview and big-picture to readers of this exciting area.

With the emergence of IoT, development of mobile sensing technology, and progress of data acquisition technology, people are not only the users and consumers of big data, but also its producers and participants. Social relation sensing, crowdsourcing, analysis of big data in SNS, and other applications closely related to human activities based on big data will be increasingly concerned and will certainly cause enormous transformations of social activities in the future society. In the future, the storage technology of big data will employ distributed databases, support transaction mechanisms similar to relational databases, and effectively handle data through grammars similar to SQL; Big data will promote the cross fusion of science; Reports, histograms, pie charts, and regression curves, etc., are frequently used to visualize results of data analysis and then new presentation forms will occur in the future.

References

[1] McGuire T, Manyika J, Chui M. Why big data is the new competitive advantage [J]. Ivey Business Journal, 2012, 76(4): 1-4.

[2] Beyer M. Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data [J]. Gartner. Archived from the original on, 2011, 10.

[3] Chen H, Chiang R H L, Storey V C. Business intelligence and analytics: From big data to big impact [J]. MIS quarterly, 2012, 36(4): 1165-1188.

[4] Gantz J, Reinsel D. Extracting value from chaos [J]. IDC iview, 2011, 1142(2011): 1-12.

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

[6] DeWitt D, Gray J. Parallel database systems: the future of high performance database systems [J]. Communications of the ACM, 1992, 35(6): 85-98.

[7] Walter T. Teradata past, present, and future [J]. UCI ISG lecture series on scalable data management, 2009.

[8] Hey T, Tansley S, Tolle K M. The fourth paradigm: data-intensive scientific discovery [M].Redmond, WA: Microsoft research, 2009.

[9] Howard J H, Kazar M L, Menees S G, et al. Scale and performance in a distributed file system [J].ACM Transactions on Computer Systems (TOCS), 1988, 6(1): 51-81.

[10] Cattell R. Scalable SQL and NoSQL data stores [J]. Acm Sigmod Record, 2011, 39(4): 12-27.

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

[12] Chaudhuri S, Dayal U, Narasayya V. An overview of business intelligence technology [J].Communications of the ACM, 2011, 54(8): 88-98.

[13] Chen C L P, Zhang C Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data [J]. Information Sciences, 2014, 275: 314-347.

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

[15] Chen M, Mao S, Liu Y. Big data: A survey [J]. Mobile Networks and Applications, 2014, 19(2): 171-209.

[16] George G, Haas M R, Pentland A. Big data and management [J]. Academy of Management Journal, 2014, 57(2): 321-326.

[17] Chen H, Chiang R H L, Storey V C. Business intelligence and analytics: From big data to big impact [J]. MIS quarterly, 2012, 36(4): 1165-1188.

[18] Wang F, Liu J. Networked wireless sensor data collection: issues, challenges, and approaches [J]. IEEE Communications Surveys & Tutorials, 2011, 13(4): 673-687.

[19] Shieh W. OFDM for flexible high-speed optical networks [J]. journal of lightwave technology, 2011, 29(10): 1560-1577.

[20] Kitchin R. The real-time city? Big data and smart urbanism [J]. GeoJournal, 2014, 79(1): 1-14.