Chapter 7
Open Issues and Outlook

Abstract In the previous chapters, we review the background and state-of-the-art of big data. In Fig. 7.1, it illustrates all the key technologies of big data introduced in this book. In this chapter, we summarize the research hot spots and suggest possible research directions of big data. We also discuss potential development trends in this broad research and application area.

7.1 Open Issues

The analysis of big data is confronted with many challenges but the current research is still in the beginning phase. Considerable research efforts are needed to improve the efficiency of data display, data storage, and data analysis.

7.1.1 Theoretical Research

Although big data is a hot research area in both academia and industry, there are many important problems remain to be solved, which are discussed below.

- *Fundamental Problems*: There is compelling need for a rigorous definition of big data, a structural model of big data, formal description of big data, and a theoretical system of data science, etc. At present, many discussions of big data look more like commercial speculation than scientific research. This is because big data is not formally and structurally defined and not strictly verified.
- *Standardization*: An evaluation system of data quality and an evaluation standard of data computing efficiency should be developed. Many solutions of big data applications claim they can improve data processing and analysis capacities in all aspects, but there is still not a unified evaluation standard and benchmark to balance the computing efficiency of big data with rigorous mathematical

M. Chen et al., *Big Data: Related Technologies, Challenges and Future Prospects*, SpringerBriefs in Computer Science, DOI 10.1007/978-3-319-06245-7_7,

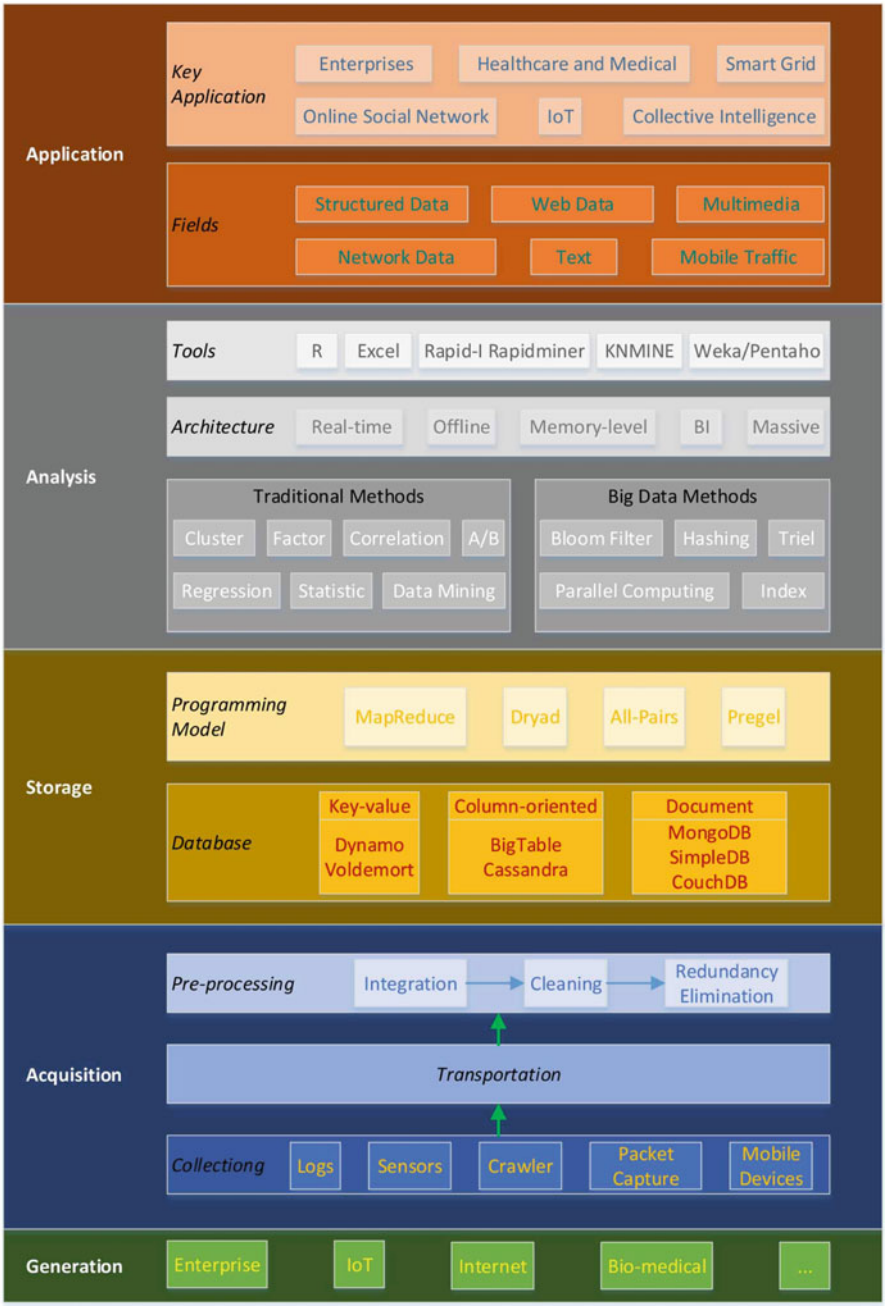

Fig. 7.1 Key technologies in big data era

methods. The performance can only be evaluated by the system is implemented and deployed, which could not horizontally compare advantages and disadvantages of various solutions and compare efficiencies before and after the use of big data. In addition, since data quality is an important basis of data preprocessing, simplification, and screening, it is also an urgent problem to effectively evaluate data quality.

- *Evolution of Big Data Computing Modes*: This includes external storage mode, data flow mode, PRAM mode, and MR mode, etc. The emergence of big data triggers the development of algorithm design, which has transformed from a computing-intensive approach into a data-intensive approach. Data transfer has been a main bottleneck of big data computing. Therefore, many new computing models tailored for big data have emerged and more such models are on the horizon.

7.1.2 Technology Development

The big data technology is still in its infancy. Many key technical problems, such as cloud computing, grid computing, stream computing, parallel computing, big data architecture, big data model, and software systems supporting big data, etc. should be fully investigated.

- *Format Conversion*: Due to wide and various data sources, heterogeneity is always a characteristic of big data, as well as a key factor which restricts the efficiency of data format conversion. If such format conversion can be made more efficient, the application of big data may create more values.
- *Big Data Transfer*: Big data transfer involves big data generation, acquisition, transmission, storage, and other data transformations in the spatial domain. As discussed, big data transfer usually incurs high costs, which is also the bottleneck for big data computing. However, data transfer is inevitable in big data applications. Improving the transfer efficiency of big data is a key factor to improve big data computing.
- *Real-time Performance*: The real-time performance of big data is also a core problem in many different application scenarios. Ways to define the life cycle of data, compute the rate of depreciation of data, and build computing models of real-time applications and online applications, will influence the values and analytical and feedback results of big data.

As big data research is advanced, new problems on big data processing arise from the traditional simple data analysis, including: (a) data re-utilization, since big data features big value but low density, with the increase of data scale, more values may be mined from re-utilization of existing data; (b) data re-organization, datasets in different businesses can be re-organized, with the total re-organized data values larger than the total datasets' value; (c) data exhaust, unstructured information or data that is a by-product of the online activities of Internet users. In big data, not

only correct data should be utilized, but also the wrong data should be utilized to generate more value. Collecting and analyzing data exhaust can provide valuable insight into the purchasing habits of consumers.

7.1.3 Practical Implications

Although there are already many successful big data applications, many practical problems should be solved:

- *Big Data Management*: the emergence of big data brings about new challenges to traditional data management. At present, many research efforts are being made on consider big data oriented database and Internet technologies, management of storage models and databases of new hardware, heterogeneous and multi-structured data integration, data management of mobile and pervasive computing, data management of SNS, and distributed data management.
- *Searching, Mining, and Analysis of Big Data*: data processing is always a research hotspot in the big data field, e.g., searching and mining of SNS models, big data searching algorithms, distributed searching, P2P searching, visualized analysis of big data, massive recommendation systems, social media systems, real-time big data mining, image mining, text mining, semantic mining, multi-structured data mining, and machine learning, etc.
- *Integration and Provenance of Big Data*: As discussed, the value acquired from a comprehensive utilization of multiple datasets is higher than the total value of individual datasets. Therefore, the integration of different data sources is a timely problem to be solved. Data integration is to integrate different datasets from different sources, which are confronted with many challenges, such as different data patterns and large amount of redundant data. Data provenance is to describe the process of data generation and evolution over time. In the big data era, data provenance is mainly used to investigate multiple datasets other than a single dataset. Therefore, it is worth of study on how to integrate data provenance information featuring different standards and from different datasets.
- *Big Data Application*: at present, the application of big data is just beginning and we shall explore and more efficiently ways to fully utilize big data. Therefore, big data applications in science, engineering, medicine, medical care, finance, business, law enforcement, education, transportation, retail, and telecommunication, big data applications in small and medium-sized businesses, big data applications in government departments, big data services, and human-computer interaction of big data, etc. are all important research problems.

7.1.4 Data Security

In IT, safety and privacy are always two key concerns. In the big data era, as data volume is fast growing, there are more severe safety risks, while the traditional data

protection methods have already been shown not applicable to big data. In particular, big data safety is confronted with the following security related challenges.

- *Big Data Privacy*: In the big data era, data privacy includes two aspects: (a) the protection of personal privacy, as the advances on data acquisition is made, personal interests, habits, and body properties, etc. of users may be more easily acquired, of which the user may not be aware. (b) Personal privacy data may also be leaked during storage, transmission, and usage, even if acquired with the permission of users. Facebook is deemed as a big data company with the most SNS data currently. Organizations that own big data usually attempt to mine valuable information in the data with advanced algorithms. The privacy data protection technology therefore is of great importance. According to a report [1], Ron Bowes, a researcher of Skull Security, acquired data in the public pages of Facebook users who fail to modify their privacy setting using an information acquisition tool. Ron Bowes packaged such data into a 2.8 GB package and created a BT seed for others to download. The analysis capacity of big data may lead to privacy mining from seemingly simple information. Therefore, privacy protection in the big data era will become a new and challenging problem.
- *Data Quality*: Data quality influences big data utilization. Low quality data wastes transmission and storage resources, and may not be usable. There are a lot of factors that may restrict data quality, for example, generation, acquisition, transmission, and transmission may all influence data quality. Data quality is mainly manifested in its accuracy, completeness, redundancy, and consistency. Even though a lot of measures have been taken to improve data quality, the quality related problems could not be completely solved. Therefore, effective methods to automatically detect data quality and repair some damaged data need to be investigated.
- *Big Data Safety Mechanism*: Big data brings challenges to data encryption due to its large scale and high variety. The performance of previous encryption methods on small and medium-scale data could not meet the demands of big data; efficient big data cryptography approaches shall be developed. Effective schemes for safety management, access control, and safety communications shall be investigated for structured, semi-structured, and unstructured data. In addition, under the multi-tenant mode, isolation, confidentiality, completeness, availability, controllability, and traceability of tenants' data should be enabled on the premise of efficiency assurance.
- *Big Data Application in Information Security*: Big data not only brings challenges to information security, but also offers new opportunities for the development of information security mechanisms. For example, we may discover potential safety loopholes and APT (Advanced Persistent Threat) after the analysis of the big data in the form of log files of an Intrusion Detection System. In addition, virus characteristics, loophole characteristics, and attack characteristics, etc. may also be more easily identified through the analysis of big data.

To sum up, the safety of big data has drawn great attention of researchers. However, there is only limited research on the representation of multi-source heterogeneous big data, measurement and semantic comprehension methods, modeling theories and computing models, distributed storage of energy efficiency optimization, and processed hardware and software system architectures, etc. Particularly, big data security, including big data credibility, big data backup and recovery technologies in various application fields, big data completeness maintenance technology, and big data security technology should be further investigated.

7.2 Outlook

The emergence of big data opens great opportunities. In the IT era, the "T" (Technology) was the main concern, while technology derives the development of data. In the big data era, with the prominence of data value and the advances in I (Information), data will drive the progress of technologies in the future. Big data will not only change the social and economic life, but also influence everyone's ways of living and thinking, which is just beginning. We could not predict the future but may take precautions for possible events to occur in the future.

- *Data With a Larger Scale, More Variety, and More Complex Structures*: Although technologies represented by Hadoop have achieved a great success, such technologies are definitely to fall behind and will be replaced given the rapid development of big data. For example, the theoretical basis of Hadoop has emerged as early as 2006. Many researchers have concerned ways to better cope with larger-scale, more various kinds of, and more complexly structured data. These efforts are represented by (Globally-Distributed Database) Spanner of Google and fault-tolerant and expandable distributed relational database F1. In the future, the storage technology of big data will be based on distributed databases, support transaction mechanisms similar to relational databases, and effectively handle data through grammars similar to SQL.
- *Data Resource Performance*: Since big data contains huge values, mastering big data means mastering resources. Through the analysis of the value chain of big data, it can be seen that its value comes from the data itself, technologies, and ideas, and the core is data resources. Without data technologies and ideas, values could not be created. The reorganization and integration of different datasets can create more values. From now on, enterprises that master big data resources may obtain huge benefits by renting and assigning the rights to use their data.
- *Big Data Promotes the Cross Fusion of Science*: Big data not only promotes the comprehensive fusion of cloud computing, Internet of Things, data center, and mobile networks, etc., but also forces the cross fusion of many disciplines. The development of big data shall explore innovative technologies and methods in big data acquisition, storage, processing, mining, and information security, etc., based on information science, and examine changes and impacts of big data on

production management, business operation and decision making, etc. of modern enterprises from the management perspective. What's more, the application of big data to specific fields needs the participation of interdisciplinary talents.

- *Visualization*: In many human-computer interaction scenarios, the principle of What You See Is What You Get is followed, e.g., text and image editors. In big data applications, mixed data may not be is very useful for decision making. Only when the analytical results are friendly displayed, it may be accepted and utilized by users. Reports, histograms, pie charts, and regression curves, etc., are frequently used to visualize results of data analysis. New presentation forms will occur in the future, e.g., Microsoft Renlifang, a social search engine, utilizes relational diagrams to express interpersonal relationship.
- *Data-Oriented*: It is well-known that programs are consisted of data structures and algorithms. In the history of program design, it is observed that the role of data is becoming increasingly more significant. In the small scale data era, in which logic is more complex than data, program design is mainly focused on processes. As business data is becoming more complex, object-oriented design methods are developed. The complexity of business data has far surpassed business logic and programs gradually transform from algorithm-intensive to data-intensive. It is anticipated data-oriented program design methods are certain to emerge, which will have far-reaching influence on the development of IT in software engineering, architecture, and model design, among others.
- *Big Data Causes the Revolution of Thinking*: In the big data era, data collection, acquisition, and analysis are more rapidly accomplished and the massive data will profoundly influence our ways of thinking. In [2], the authors summarizes the thinking revolution caused by big data as follows:
  - During data analysis, we will try to utilize all data other than only analyzing a little sample data.
  - Compared with accurate data, we would like to accept numerous and complicated data.
  - We shall pay greater attention to correlations between things other than exploring causal relationship.
  - The simple algorithms of big data are more effective than complex algorithms of small data.
  - Analytical results of big data will reduce hasty and subjective factors during decision making and data scientists will replace "experts."
- *Managing Large-scale FlowTable for Software-Defined Networking with Big Data Techniques*: In the past few years, software-defined networking (SDN) has been the buzz of the networking world. It was originally proposed to accelerate networking innovations in legacy campus networks called OpenFlow, which comprises a number of closed networking boxes with diverse functionalities (such as routing, switching, firewall, etc.) [3]. It is observed that, plenty of emerging networking problems appeared in the era when cloud computing meets big data applications, and SDN seems to be extremely suitable for solving those problems in respect of network efficiency, scalability, flexibility, agility, as well

as operation and maintenance complexity. In the specification of OpenFlow, one of the most important concept is FlowTable, which includes a large number of rules to process network packets. Obviously, it is a challenge to manage the large-scale FlowTables. A promising way is to implement SDN with big data techniques, to effectively store, process and utilize FlowTable, and increase the speed of searching rules.

- *5G Wireless Networks: Supporting Technology for Mobile Big Data*: With the emergence of cloud computing as an important information technology in support of virtualized services, it becomes promising to design 5G wireless networks by exploiting recent advances relevant to network function virtualization and benefiting from advanced virtualization techniques of cloud computing to build efficient and scalable networking infrastructures. Researchers have been designing new architectures for elastically composing and operating a virtual end-to-end network platform on demand on top of fragmented physical infrastructures provided by federated cloud. SDN techniques have been seen as promising enablers for this vision of carrier cloud, which will likely play a crucial role in the design of 5G wireless networks.

  Due to the huge explosion in mobile data of a hyperconnected society, "Can Big Data go Mobile?" now becomes a challenging problem which would be addressed by 5G technologies. Though 5G wireless provides the possibility to enable the mobility of big data, there are various research problems towards the realization of the brand-new networking system, such as 5G network architecture, SDN and network virtualization techniques for enabling 5G, resource allocation algorithms in 5G, and 5G-related control protocols and optimization techniques. In an energy efficient, flexible, connectivity-scalable and secure manner, new business models beyond IaaS, PaaS and SaaS, such as Network as a Service (NaaS), and Knowledge as a Service (KaaS), are expected to emerge. Especially, Big Data as a Service (BDaaS) or Big Data Analysis as a Service (BDAaaS) could emerge, facilitating the efficient storage and analysis for the exploding mobile data.

Throughout the history of human society, the demands and willingness of human beings are always the source powers to promote scientific and technological progress. In the big data era, big data may provides reference answers for human beings to make decisions through mining and analytical processing, but could not replace human thinking. It is human thinking that promotes the widespread utilizations of big data. Big data is more like an extendable and expandable human brain other than a substitute of human brain. With the emergence of Internet of Things, development of mobile sensing technology, and progress of data acquisition technology, people are not only the user and consumer of big data, but also its producer and participant. 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 changes of social activities in the future society.

References

1. Predrag Tasevski. Password attacks and generation strategies. *Tartu University: Faculty of Mathematics and Computer Sciences*, 2011.
2. Viktor Mayer-Schönberger and Kenneth Cukier. *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Eamon Dolan/Houghton Mifflin Harcourt, 2013.
3. McKeown, Nick and Anderson, Tom and Balakrishnan, Hari and Parulkar, Guru and Peterson, Larry and Rexford, Jennifer and Shenker, Scott and Turner, Jonathan. OpenFlow: enabling innovation in campus networks. *ACM SIGCOMM Computer Communication Review*, 38(2):69–77, 2008.