Chapter 1
Introduction

Abstract The term of *big data* was coined under the explosive increase of global data and was mainly used to describe these enormous datasets. In this chapter, we introduce the definition of big data, and review its evolution in the past 20 years. In particular, we introduce the defining features of big data, as well as its 4Vs characteristics, including Volume, Variety, Velocity, and Value. The challenges brought about by big data is also examined in this chapter.

1.1 Dawn of the Big Data Era

Over the past 20 years, data has increased in a large scale in various fields. According to a report from International Data Corporation (IDC), in 2011, the overall created and copied data volume in the world was 1.8ZB ($\approx 10^{21}$B), which has increased by nearly nine times within 5 years [1]. Such figure will double at least every other 2 years in the near future.

The term of *big data* was coined under the explosive increase of global data and was mainly used to describe these enormous datasets. Compared with traditional datasets, big data generally includes masses of unstructured data that need more real-time analysis. In addition, big data also brings new opportunities for discovering new values, helps us to gain an in-depth understanding of the hidden values, and incurs new challenges, e.g., on how to effectively organize and manage such data. At present, big data has attracted considerable interest from industry, academia, and government agencies. For example, issues on big data are often covered in public media, including *The Economist* [2, 3], *New York Times* [4], and *National Public Radio* [5, 6]. Two premier scientific journals, *Nature* and *Science*, also started special columns to discuss the importance and challenges of big data [7, 8]. Many government agencies announced major plans to accelerate big data research and applications [9], and industries also become interested in the high potential of big data recently. The era of big data is coming beyond all doubt [12].

M. Chen et al., *Big Data: Related Technologies, Challenges and Future Prospects*, SpringerBriefs in Computer Science, DOI 10.1007/978-3-319-06245-7_1,

Recently, the rapid growth of big data mainly comes from people's daily life, especially related to the service of Internet companies. For example, Google processes data of hundreds of PB and Facebook generates log data of over 10 Petabyte (PB) per month; Baidu, a Chinese company, processes data of tens of PB and Taobao, a subsidiary of Alibaba, generates data of tens of Terabyte (TB) on online trading per day. While the amount of large datasets is drastically rising, it also brings about many challenging problems demanding prompt solutions. First, the latest advances of information technology (IT) make it more easily to generate data. For example, on average, 72 h of videos are uploaded to YouTube in every minute [13]. Therefore, we are confronted with the main challenge of collecting and integrating massive data from widely distributed data sources. Second, the collected data is increasingly growing, which causes a problem of how to store and manage such huge, heterogeneous datasets with moderate requirements on hardware and software infrastructure. Third, in consideration of the heterogeneity, scalability, real-time, complexity, and privacy of big data, we shall effectively "mine" the datasets at different levels with analysis, modeling, visualization, forecast, and optimization techniques, so as to reveal its intrinsic property and improve decision making.

The rapid growth of cloud computing and the Internet of Things (IoT) further promote the sharp growth of data. Cloud computing provides safeguarding, access sites, and channels for data asset. In the paradigm of IoT, sensors all over the world are collecting and transmitting data which will be stored and processed in the cloud. Such data in both quantity and mutual relations will far surpass the capacities of the IT architectures and infrastructure of existing enterprises, and its realtime requirement will greatly stress the available computing capacity. Figure 1.1 illustrates the boom of the global data volume.

1.2 Definition and Features of Big Data

Big data is an abstract concept. Apart from masses of data, it also has some other features, which determine the difference between itself and "massive data" or "very big data." At present, although the importance of big data has been generally recognized, people still have different opinions on its definition. In general, big data refers to 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.

In 2010, Apache Hadoop defined big data as "datasets which could not be captured, managed, and processed by general computers within an acceptable scope." On the basis of this definition, in May 2011, McKinsey & Company, a global consulting agency announced Big Data as "the Next Frontier for Innovation, Competition, and Productivity." Big data shall mean such datasets which could

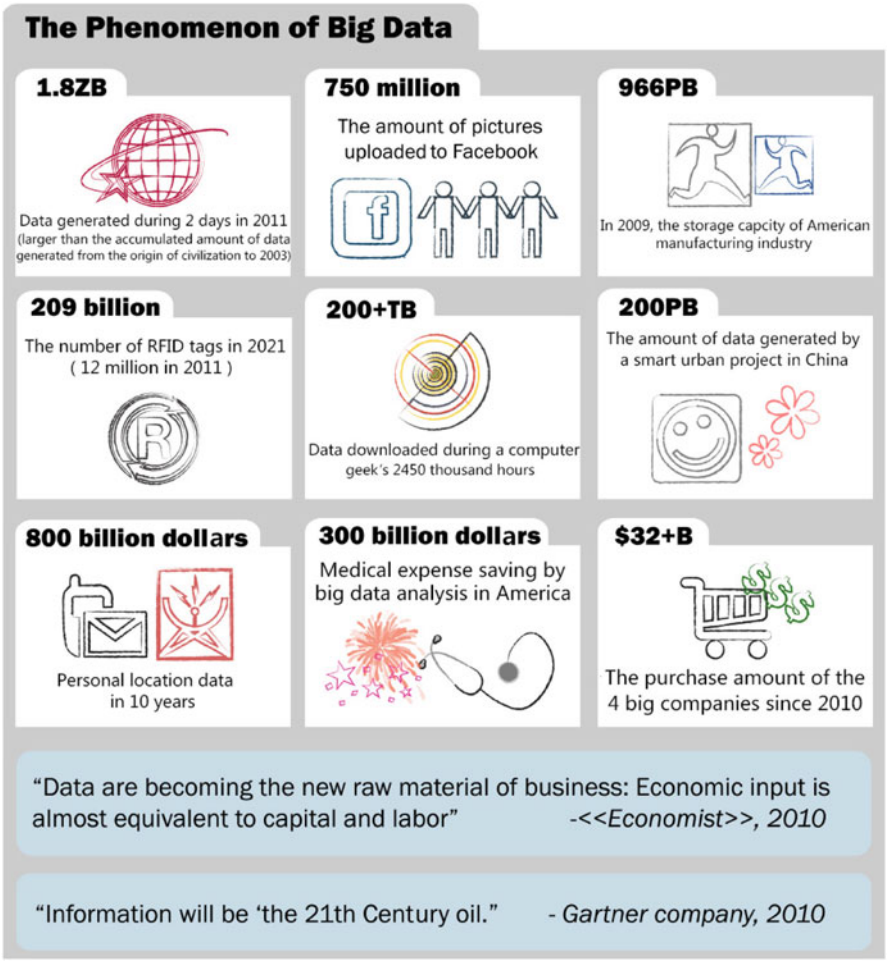

Fig. 1.1 Illustrating the continuously increasing big data

not be acquired, stored, and managed by classic database software. This definition includes two connotations: First, the dataset volumes that conform to the standard of big data are changing, and may grow over time or with technological advances; Second, the dataset volumes that conform to the standard of big data in different applications differ from each other. At present, big data generally range from several TB to several PB [12]. 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.

As a matter of fact, big data has been defined as early as 2001. Doug Laney, an analyst of META (presently Gartner) defined challenges and opportunities brought

about by the increased data with a 3Vs model, i.e., the increase of Volume, Velocity, and Variety, in a research report [14]. Although such a model was not originally used to define big data, Gartner and many other enterprises, including IBM [15] and some research departments of Microsoft [16] still used the "3Vs" model to describe big data within the following 10 years [17]. In the "3Vs" model, Volume means, with the generation and collection of massive data, data scale becomes increasingly huge; Velocity means the timeliness of big data, specifically, data collection and analysis, etc., must be rapidly and timely conducted, so as to maximumly utilize the commercial value of big data; Variety indicates the various types of data, which include semi-structured and unstructured data such as audio, video, webpage, and text, as well as traditional structured data.

However, others have different opinions, including IDC, one of the most influential leaders in big data and its research fields. 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" [1]. With this definition, characteristics of big data can 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 Fig. 1.2. 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" [13].

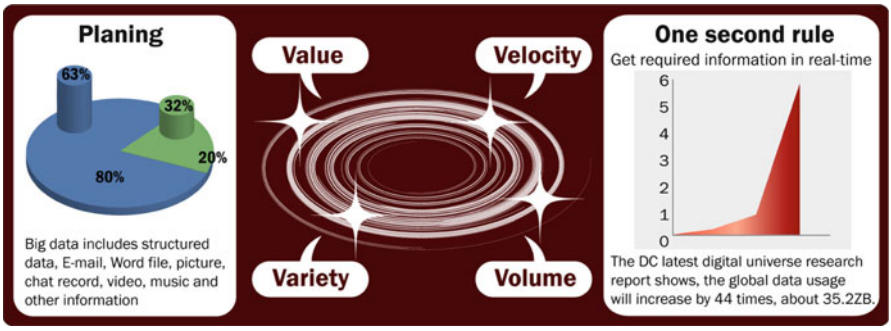

Fig. 1.2 The 4Vs feature of big data

In addition, the US National Institute of Standards and Technology (NIST) defines big data as "Big data shall mean the data of which the data volume, acquisition speed, or data representation limits the capacity of using traditional relational methods to conduct effective analysis or the data which may be effectively processed with important horizontal zoom technologies," which focuses on the

technological aspect of big data. It indicates that efficient methods or technologies need to be developed and used to analyze and process big data.

There have been considerable discussions from both industry and academia on the definition of big data [10, 11]. In addition to developing a proper definition, the big data research should also focus on how to extract its value, how to make use of data, and how to transform"a bunch of data" into "big data."

1.3 Big Data Value

McKinsey & Company observed how big data created values after in-depth research on the U.S. healthcare, the EU public sector administration, the U.S. retail, the global manufacturing, and the global personal location data. Through research on the five core industries that represent the global economy, the McKinsey report pointed out that big data may give a full play to the economic function, improve the productivity and competitiveness of enterprises and public sectors, and create huge benefits for consumers. In [12], McKinsey summarized the values that big data could create: if big data could be creatively and effectively utilized to improve efficiency and quality, the potential value of the U.S. medical industry gained through data may surpass USD 300 billion, thus reducing the U.S. healthcare expenditure by over 8 %; retailers that fully utilize big data may improve their profit by more than 60 %; big data may also be utilized to improve the efficiency of government operations, such that the developed economies in Europe could save over EUR 100 billion (which excludes the effect of reduced frauds, errors, and tax difference).

The McKinsey report is regarded as prospective and predictive, while the following facts may validate the values of big data. During the 2009 flu pandemic, Google obtained timely information by analyzing big data, which even provided more valuable information than that provided by disease prevention centers. Nearly all countries required hospitals inform agencies such as disease prevention centers of new type of influenza cases. However, patients usually did not see doctors immediately when they got infected. It also took some time to send information from hospitals to disease prevention centers, and for disease prevention centers to analyze and summarize such information. Therefore, when the public is aware of the pandemic of a new type of influenza, the disease may have already spread for one to two weeks with a serious hysteretic nature. Google found that during the spreading of influenza, entries frequently sought at its search engines would be different from those at ordinary times, and the usage frequencies of the entries were correlated to the influenza spreading in both time and location. Google found 45 search entry groups that were closely relevant to the outbreak of influenza and incorporated them in specific mathematic models to forecast the spreading of influenza and even to predict places where influenza will spread from. The related research results have been published in Nature [18].

In 2008, Microsoft purchased Farecast, a sci-tech venture company in the U.S. Forecast has an airline ticket forecasting system that predicts the trends and rising/dropping ranges of airline ticket prices. The system has been incorporated into the Bing search engine of Microsoft. By 2012, the system has saved nearly USD 50 per ticket per passenger, with the forecast accuracy as high as 75 %.

At present, data has become an important production factor that could be comparable to material assets and human capital. As multimedia, social media, and IoT are fast evolving, enterprises will collect more information, leading to an exponential growth of data volume. Big data will have a huge and increasing potential in creating values for businesses and consumers.

1.4 The Development of Big Data

In late 1970s, the concept of "database machine" emerged, which is a technology specially used for storing and analyzing data. With the increase of data volume, the storage and processing capacity of a single mainframe computer system has become inadequate. In the 1980s, people proposed "share nothing," a parallel database system, to meet the demand of the increasing data volume [19]. The share nothing system architecture is based on the use of cluster and every machine has its own processor, storage, and disk. Teradata system was the first successful commercial parallel database system. Such database became very popular lately. On June 2, 1986, a milestone event occurred, when Teradata delivered the first parallel database system with a storage capacity of 1TB to Kmart to help the large-scale retail company in North America to expand its data warehouse [20]. In late 1990s, the advantages of the parallel database was widely recognized in the database field.

However, many challenges on big data arose. With the development of Internet services, indexes and queried contents were rapidly growing. Therefore, search engine companies had to face the challenges of handling such big data. Google created GFS [21] and MapReduce [22] programming models to cope with the challenges brought about by data management and analysis at the Internet scale. In addition, contents generated by users, sensors, and other ubiquitous data sources also drive the overwhelming data flows, which required a fundamental change on the computing architecture and large-scale data processing mechanism. In January 2007, Jim Gray, a pioneer of database software, called such transformation "The Fourth Paradigm" [23]. He also thought the only way to cope with such a paradigm was to develop a new generation of computing tools to manage, visualize, and analyze massive data. In June 2011, another milestone event occurred, when EMC/IDC published a research report titled *Extracting Values from Chaos* [1], which introduced the concept and potential of big data for the first time. This research report aroused 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. Taking IBM as an example, since 2005, IBM has invested USD 16 billion on 30 acquisitions related to big data. In academia, big data was also under the spotlight. In 2008, Nature published the 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 Initiative*, which was a second major scientific and technological development initiative after the *Information Highway Initiative* in 1993. In July 2012, the *Japan's ICT* project issued by 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.

1.5 Challenges of Big Data

The sharply increasing data deluge in the big data era brings huge challenges on data acquisition, storage, management and analysis. Traditional data management and analytics 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 cannot 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 [24] and NoSQL [25] 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 analytics systems.

Some literatures [26–28] discuss obstacles to be overcome in the development of big data applications. Some key challenges are listed as follows:

- *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.
- *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.
- *Data Life Cycle Management*: compared with the relatively slow advances of storage systems, pervasive sensors 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, an importance principle related to the analytical value should be developed to decide which data shall be stored and which data shall be discarded.
- *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. Even so, there are still some problems of non-relational databases in their performance and particular applications. We shall find a compromising solution between RDBMSs and non-relational databases. For example, some enterprises have utilized a mixed database architecture that integrates the advantages of both types of database (e.g., Facebook and Taobao). More research is needed on the in-memory database and sample data based on approximate analysis.
- *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 the 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 the sensitive data, to ensure its safety.

- *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 mechanisms shall be established for big data while expandability and accessibility are both ensured.
- *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.
- *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.

References

1. John Gantz and David Reinsel. Extracting value from chaos. *IDC iView*, pages 1–12, 2011.
2. Kenneth Cukier. *Data, data everywhere: A special report on managing information*. Economist Newspaper, 2010.
3. Drowning in numbers - digital data will flood the planet- and help us understand it better. http://www.economist.com/blogs/dailychart/2011/11/bigdata-0, 2011.
4. Steve Lohr. The age of big data. *New York Times*, 11, 2012.
5. Noguchi Yuki. Following digital breadcrumbs to big data gold. http://www.npr.org/2011/11/29/142521910/thedigitalbreadcrumbs-that-lead-to-big-data, 2011.
6. Noguchi Yuki. The search for analysts to make sense of big data. http://www.npr.org/2011/11/30/142893065/the-searchforanalysts-to-make-sense-of-big-data, 2011.
7. Big data. http://www.nature.com/news/specials/bigdata/index.html, 2008.
8. Special online collection: Dealing with big data. http://www.sciencemag.org/site/special/data/, 2011.
9. Fact sheet: Big data across the federal government. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_3_29_2012.pdf, 2012.
10. O. R. Team. *Big Data Now: Current Perspectives from O'Reilly Radar*. O'Reilly Media, 2011.
11. M Grobelnik. Big data tutorial. http://videolectures.net/eswc2012grobelnikbigdata/, 2012.
12. James Manyika, McKinsey Global Institute, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute, 2011.
13. 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.
14. Douglas Laney. 3-d data management: Controlling data volume, velocity and variety. *META Group Research Note, February*, 6, 2001.
15. Paul Zikopoulos, Chris Eaton, et al. *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media, 2011.
16. Erik Meijer. The world according to linq. *Communications of the ACM*, 54(10):45–51, 2011.
17. Mark Beyer. Gartner says solving 'big data' challenge involves more than just managing volumes of data. *Gartner*. http://www.gartner.com/it/page.jsp, 2011.

18. Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.
19. David DeWitt and Jim Gray. Parallel database systems: the future of high performance database systems. *Communications of the ACM*, 35(6):85–98, 1992.
20. T Walter. Teradata past, present, and future. UCI ISG Lecture Series on Scalable Data Management.
21. Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In *ACM SIGOPS Operating Systems Review*, volume 37, pages 29–43. ACM, 2003.
22. Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.
23. Anthony JG Hey, Stewart Tansley, Kristin Michele Tolle, et al. The fourth paradigm: data-intensive scientific discovery. 2009.
24. John H Howard, Michael L Kazar, Sherri G Menees, David A Nichols, Mahadev Satyanarayanan, Robert N Sidebotham, and Michael J West. Scale and performance in a distributed file system. *ACM Transactions on Computer Systems (TOCS)*, 6(1):51–81, 1988.
25. Rick Cattell. Scalable sql and nosql data stores. *ACM SIGMOD Record*, 39(4):12–27, 2011.
26. Alexandros Labrinidis and HV Jagadish. Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, 5(12):2032–2033, 2012.
27. Surajit Chaudhuri, Umeshwar Dayal, and Vivek Narasayya. An overview of business intelligence technology. *Communications of the ACM*, 54(8):88–98, 2011.
28. D Agrawal, P Bernstein, E Bertino, S Davidson, U Dayal, M Franklin, J Gehrke, L Haas, A Halevy, J Han, et al. Challenges and opportunities with big data. a community white paper developed by leading researchers across the united states, 2012.