Mobile Netw Appl (2014) 19:171–209
DOI 10.1007/s11036-013-0489-0

Big Data: A Survey

Min Chen · Shiwen Mao · Yunhao Liu

Abstract In this paper, we review the background and state-of-the-art of big data. We first introduce the general background of big data and review related technologies, such as could computing, Internet of Things, data centers, and Hadoop. We then focus on the four phases of the value chain of big data, i.e., data generation, data acquisition, data storage, and data analysis. For each phase, we introduce the general background, discuss the technical challenges, and review the latest advances. We finally examine the several representative applications of big data, including enterprise management, Internet of Things, online social networks, medial applications, collective intelligence, and smart grid. These discussions aim to provide a comprehensive overview and big-picture to readers of this exciting area. This survey is concluded with a discussion of open problems and future directions.

Keywords Big data · Cloud computing · Internet of things · Data center · Hadoop · Smart grid · Big data analysis

M. Chen (✉)
School of Computer Science and Technology,
Huazhong University of Science and Technology,
1037 Luoyu Road, Wuhan, 430074, China
e-mail: minchen2012@hust.edu.cn; minchen@ieee.org

S. Mao
Department of Electrical & Computer Engineering,
Auburn University, 200 Broun Hall, Auburn,
AL 36849-5201, USA
e-mail: smao@ieee.org

Y. Liu
TNLIST, School of Software, Tsinghua University, Beijing, China
e-mail: yunhao@greenorbs.com

1 Background

1.1 Dawn of 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 increased by nearly nine times within five years [1]. This figure will double at least every other two years in the near future.

Under the explosive increase of global data, the term of big data is mainly used to describe enormous datasets. Compared with traditional datasets, big data typically includes masses of unstructured data that need more real-time analysis. In addition, big data also brings about new opportunities for discovering new values, helps us to gain an in-depth understanding of the hidden values, and also incurs new challenges, e.g., how to effectively organize and manage such datasets.

Recently, industries become interested in the high potential of big data, and many government agencies announced major plans to accelerate big data research and applications [2]. In addition, issues on big data are often covered in public media, such as *The Economist* [3, 4], *New York Times* [5], and *National Public Radio* [6, 7]. Two premier scientific journals, *Nature* and *Science*, also opened special columns to discuss the challenges and impacts of big data [8, 9]. The era of big data has come beyond all doubt [10].

Nowadays, big data related to the service of Internet companies grow rapidly. For example, Google processes data of hundreds of Petabyte (PB), Facebook generates log data of over 10 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) for online trading per day. Figure 1 illustrates the boom of the global data volume. While the amount of large datasets is drastically rising, it also brings about many challenging problems demanding prompt solutions:

- The latest advances of information technology (IT) make it more easily to generate data. For example, on average, 72 hours of videos are uploaded to YouTube in every minute [11]. Therefore, we are confronted with the main challenge of collecting and integrating massive data from widely distributed data sources.
- 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 to 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 also greatly stress the available computing capacity. The increasingly growing data cause a problem of how to store and manage such huge heterogeneous datasets with moderate requirements on hardware and software infrastructure.
- In consideration of the heterogeneity, scalability, realtime, complexity, and privacy of big data, we shall effectively "mine" the datasets at different levels during the analysis, modeling, visualization, and forecasting, so as to reveal its intrinsic property and improve the decision making.

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."

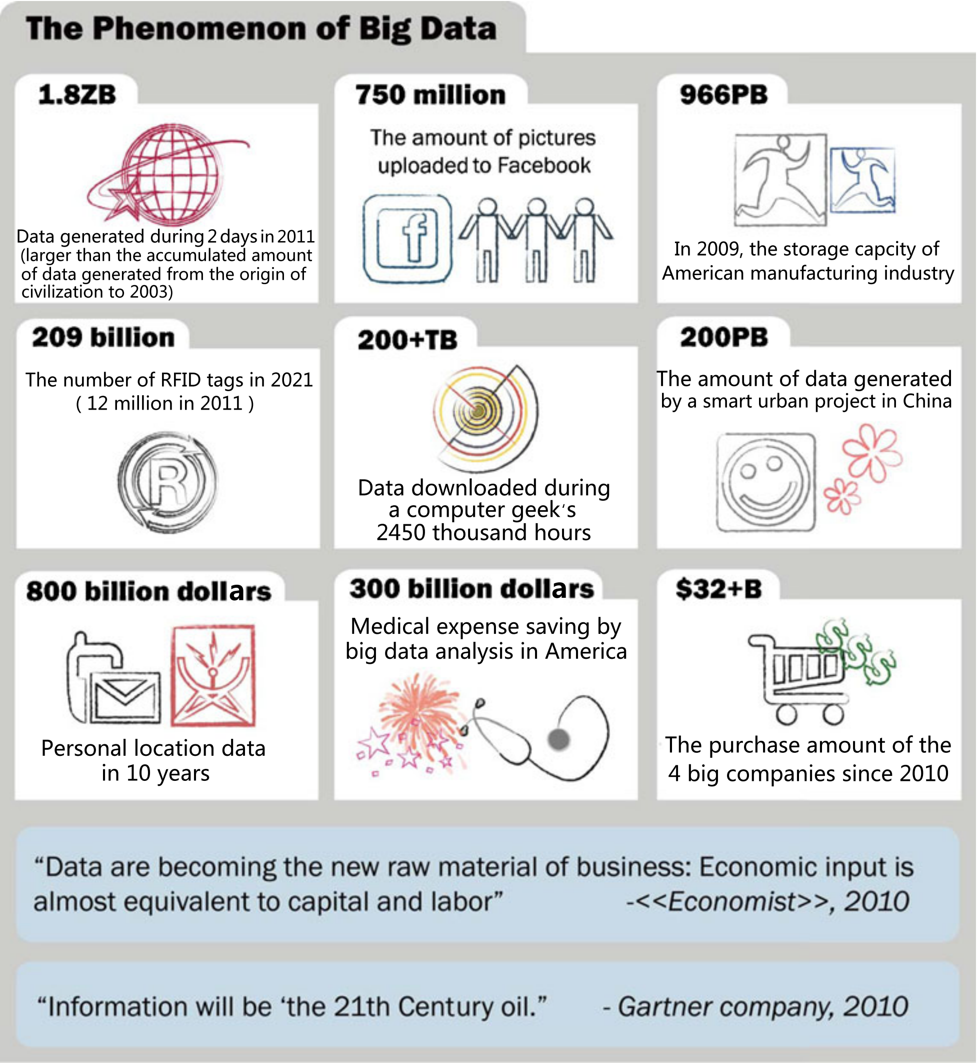

The Phenomenon of Big Data
1.8ZB
Data generated during 2 days in 2011 (larger than the accumulated amount of data generated from the origin of civilization to 2003)
750 million
The amount of pictures uploaded to Facebook
966PB
In 2009, the storage capcity of American manufacturing industry
209 billion
The number of RFID tags in 2021 (12 million in 2011)
200+TB
Data downloaded during a computer geek's 2450 thousand hours
200PB
The amount of data generated by a smart urban project in China
800 billion dollars
Personal location data in 10 years
300 billion dollars
Medical expense saving by big data analysis in America
$32+B
The purchase amount of the 4 big companies since 2010
"Data are becoming the new raw material of business: Economic input is almost equivalent to capital and labor" -<<Economist>>, 2010
"Information will be 'the 21th Century oil." - Gartner company, 2010

Fig. 1 The continuously increasing 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 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.

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 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 [10]. 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 increased data with a 3Vs model, i.e., the increase of Volume, Velocity, and Variety, in a research report [12]. Although such a model was not originally used to define big data, Gartner and many other enterprises, including IBM [13] and some research departments of Microsoft [14] still used the "3Vs" model to describe big data within the following ten years [15]. In the "3Vs" model, Volume means, with the generation and collection of masses of data, data scale becomes increasingly big; 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 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 Fig. 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." [11]

In addition, 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 [16, 17]. 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."

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 [10], 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 expenditure for the U.S. healthcare 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).

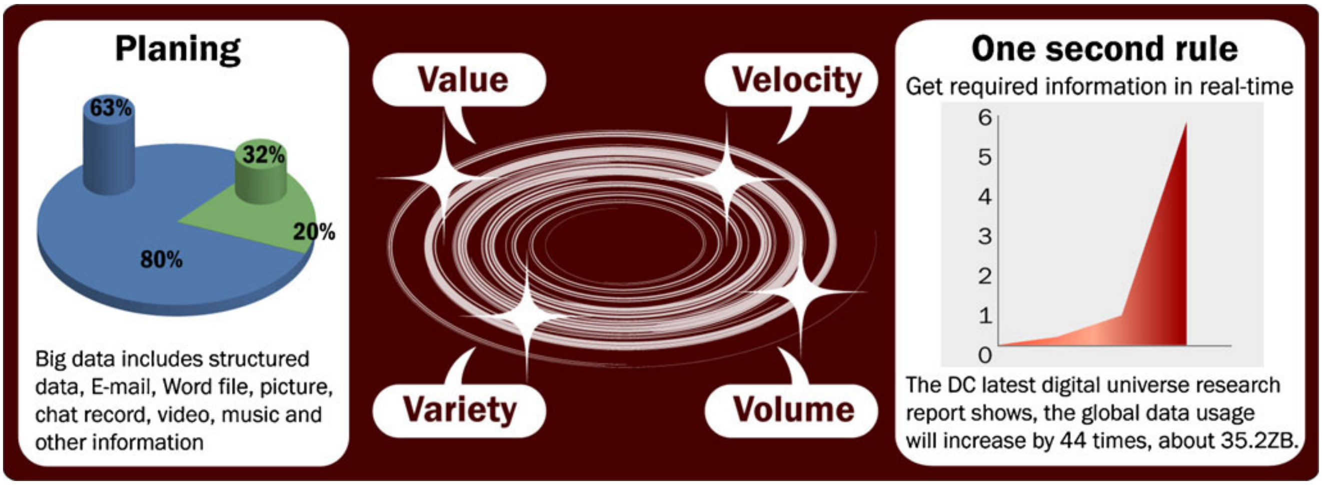

Planing
63%
32%
20%
80%
Big data includes structured data, E-mail, Word file, picture, chat record, video, music and other information
Value
Velocity
Variety
Volume
One second rule
Get required information in real-time
6
5
4
3
2
1
0
The DC latest digital universe research report shows, the global data usage will increase by 44 times, about 35.2ZB.

Fig. 2 The 4Vs feature of big data

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 the 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 the new type of influenza, the disease may have already spread for one to two weeks with a 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 use 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 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. Farecast has an airline ticket forecast system that predicts the trends and rising/dropping ranges of airline ticket price. 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 forecasted 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 developing, 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 the 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 became 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 the storage capacity of 1TB to Kmart to help the large-scale retail company in North America to expand its data warehouse [20]. In the late 1990s, the advantages of parallel database was widely recognized in the database field.

However, many challenges on big data arose. With the development of Internet servies, 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 feuled 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 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* [1], 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. 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 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.

1.5 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 [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 analysis systems.

Some literature [26–28] discuss obstacles in the development of big data applications. The 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 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.
- *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 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.
- *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.
- *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.

2 Related technologies

In order to gain a deep understanding of big data, this section will introduce several fundamental technologies that are closely related to big data, including cloud computing, IoT, data center, and Hadoop.

2.1 Relationship between cloud computing and big data

Cloud computing is closely related to big data. The key components of cloud computing are shown in Fig. 3. Big data is the object of the computation-intensive operation and stresses the storage capacity of a cloud system. The main objective of cloud computing is to use huge computing and storage resources under concentrated management, so as to provide big data applications with fine-grained computing capacity. The development of cloud computing provides solutions for the storage and processing of big data. On the other hand, the emergence of big data also accelerates the development of cloud computing. The distributed storage technology based on cloud computing can effectively manage big data; the parallel computing capacity by virtue of cloud computing can improve the efficiency of acquisition and analyzing big data.

Even though there are many overlapped technologies in cloud computing and big data, they differ in the following two aspects. First, the concepts are different to a certain extent. Cloud computing transforms the IT architecture while big data influences business decision-making. However, big data depends on cloud computing as the fundamental infrastructure for smooth operation.

Second, big data and cloud computing have different target customers. Cloud computing is a technology and product targeting Chief Information Officers (CIO) as an advanced IT solution. Big data is a product targeting Chief Executive Officers (CEO) focusing on business operations. Since the decision makers may directly feel the pressure from market competition, they must defeat business opponents in more competitive ways. With the advances of big data and cloud computing, these two technologies are certainly and increasingly entwine with each other. Cloud computing, with functions similar to those of computers and operating systems, provides system-level resources; big data

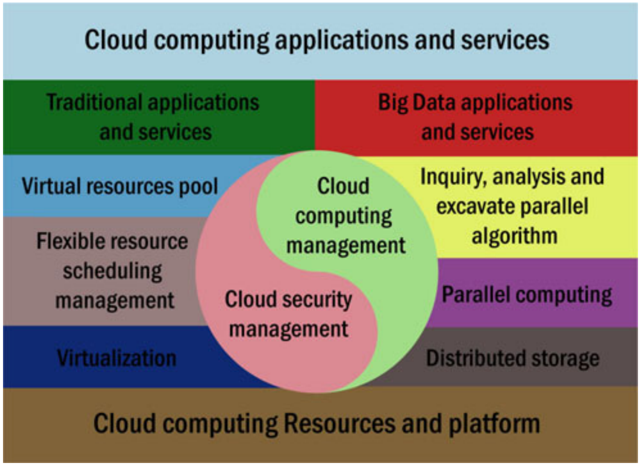

Cloud computing applications and services
Traditional applications and services
Big Data applications and services
Virtual resources pool
Cloud computing management
Inquiry, analysis and excavate parallel algorithm
Flexible resource scheduling management
Cloud security management
Parallel computing
Virtualization
Distributed storage
Cloud computing Resources and platform

Fig. 3 Key components of cloud computing

operates in the upper level supported by cloud computing and provides functions similar to those of database and efficient data processing capacity. Kissinger, President of EMC, indicated that the application of big data must be based on cloud computing.

The evolution of big data was driven by the rapid growth of application demands and cloud computing developed from virtualized technologies. Therefore, cloud computing not only provides computation and processing for big data, but also itself is a service mode. To a certain extent, the advances of cloud computing also promote the development of big data, both of which supplement each other.

2.2 Relationship between IoT and big data

In the IoT paradigm, an enormous amount of networking sensors are embedded into various devices and machines in the real world. Such sensors deployed in different fields may collect various kinds of data, such as environmental data, geographical data, astronomical data, and logistic data. Mobile equipments, transportation facilities, public facilities, and home appliances could all be data acquisition equipments in IoT, as illustrated in Fig. 4.

The big data generated by IoT has different characteristics compared with general big data because of the different types of data collected, of which the most classical characteristics include heterogeneity, variety, unstructured feature, noise, and high redundancy. Although the current IoT data is not the dominant part of big data, by 2030, the quantity of sensors will reach one trillion and then the IoT data will be the most important part of big data, according to the forecast of HP. A report from Intel pointed out that big data in IoT has three features that conform to the big data paradigm: (i) abundant terminals generating masses of data; (ii) data generated by IoT is usually semi-structured or unstructured; (iii) data of IoT is useful only when it is analyzed.

At present, the data processing capacity of IoT has fallen behind the collected data and it is extremely urgent to accelerate the introduction of big data technologies to promote the development of IoT. Many operators of IoT realize the importance of big data since the success of IoT is hinged upon the effective integration of big data and cloud computing. The widespread deployment of IoT will also bring many cities into the big data era.

There is a compelling need to adopt big data for IoT applications, while the development of big data is already legged behind. It has been widely recognized that these two technologies are inter-dependent and should be jointly developed: on one hand, the widespread deployment of IoT drives the high growth of data both in quantity and category, thus providing the opportunity for the application and development of big data; on the other hand, the application of big data technology to IoT also accelerates the research advances and business models of of IoT.

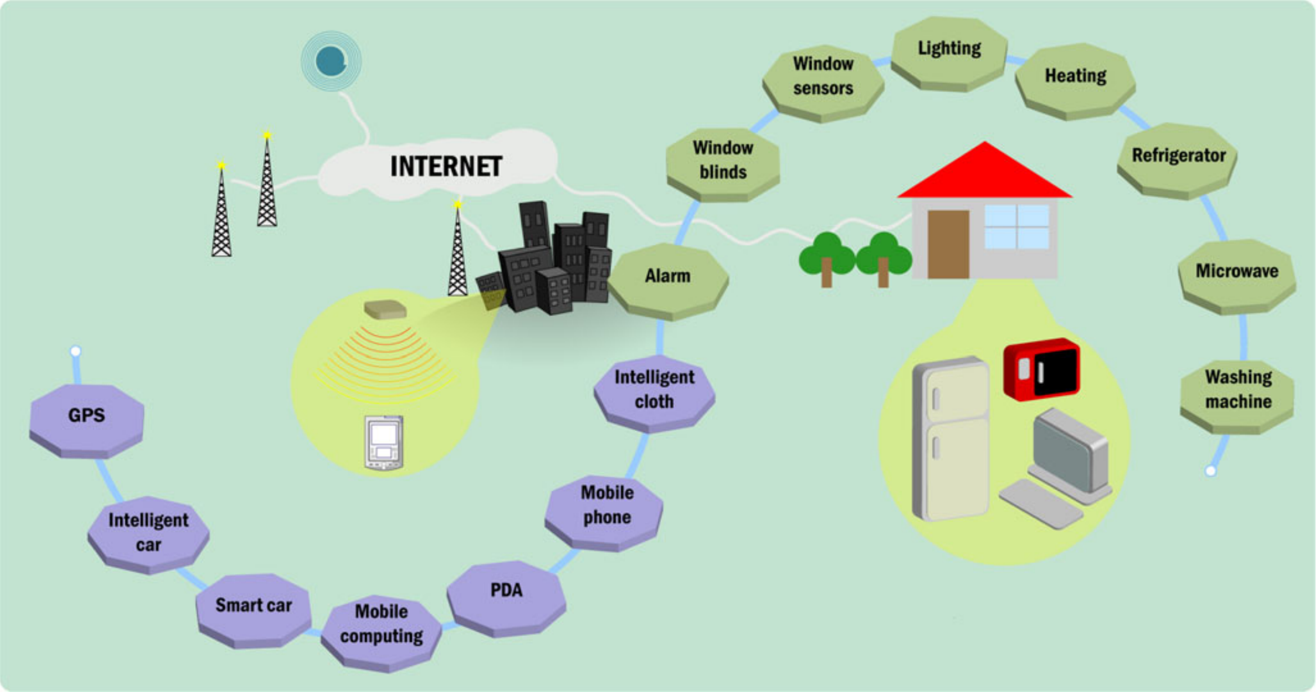


Fig. 4 Illustration of data acquisition equipment in IoT

2.3 Data center

In the big data paradigm, the data center not only is a platform for concentrated storage of data, but also undertakes more responsibilities, such as acquiring data, managing data, organizing data, and leveraging the data values and functions. Data centers mainly concern “data” other than “center.” It has masses of data and organizes and manages data according to its core objective and development path, which is more valuable than owning a good site and resource. The emergence of big data brings about sound development opportunities and great challenges to data centers. Big data is an emerging paradigm, which will promote the explosive growth of the infrastructure and related software of data center. The physical data center network is the core for supporting big data, but, at present, is the key infrastructure that is most urgently required [29].

- Big data requires data center provide powerful backstage support. The big data paradigm has more stringent requirements on storage capacity and processing capacity, as well as network transmission capacity. Enterprises must take the development of data centers into consideration to improve the capacity of rapidly and effectively processing of big data under limited price/performance ratio. The data center shall provide the infrastructure with a large number of nodes, build a high-speed internal network, effectively dissipate heat, and effective backup data. Only when a highly energy-efficient, stable, safe, expandable, and redundant data center is built, the normal operation of big data applications may be ensured.
- The growth of big data applications accelerates the revolution and innovation of data centers. Many big data applications have developed their unique architectures and directly promote the development of storage, network, and computing technologies related to data center. With the continued growth of the volumes of structured and unstructured data, and the variety of sources of analytical data, the data processing and computing capacities of the data center shall be greatly enhanced. In addition, as the scale of data center is increasingly expanding, it is also an important issue on how to reduce the operational cost for the development of data centers.
- Big data endows more functions to the data center. In the big data paradigm, data center shall not only concern with hardware facilities but also strengthen soft capacities, i.e., the capacities of acquisition, processing, organization, analysis, and application of big data. The data center may help business personnel analyze the existing data, discover problems in business operation, and develop solutions from big data.

2.4 Relationship between hadoop and big data

Presently, Hadoop is widely used in big data applications in the industry, e.g., spam filtering, network searching, clickstream analysis, and social recommendation. In addition, considerable academic research is now based on Hadoop. Some representative cases are given below. As declared in June 2012, Yahoo runs Hadoop in 42,000 servers at four data centers to support its products and services, e.g., searching and spam filtering, etc. At present, the biggest Hadoop cluster has 4,000 nodes, but the number of nodes will be increased to 10,000 with the release of Hadoop 2.0. In the same month, Facebook announced that their Hadoop cluster can process 100 PB data, which grew by 0.5 PB per day as in November 2012. Some well-known agencies that use Hadoop to conduct distributed computation are listed in [30]. In addition, many companies provide Hadoop commercial execution and/or support, including Cloudera, IBM, MapR, EMC, and Oracle.

Among modern industrial machinery and systems, sensors are widely deployed to collect information for environment monitoring and failure forecasting, etc. Bahga and others in [31] proposed a framework for data organization and cloud computing infrastructure, termed CloudView. CloudView uses mixed architectures, local nodes, and remote clusters based on Hadoop to analyze machine-generated data. Local nodes are used for the forecast of real-time failures; clusters based on Hadoop are used for complex offline analysis, e.g., case-driven data analysis.

The exponential growth of the genome data and the sharp drop of sequencing cost transform bio-science and biomedicine to data-driven science. Gunarathne et al. in [32] utilized cloud computing infrastructures, Amazon AWS, Microsoft Azune, and data processing framework based on MapReduce, Hadoop, and Microsoft DryadLINQ to run two parallel bio-medicine applications: (i) assembly of genome segments; (ii) dimension reduction in the analysis of chemical structure. In the subsequent application, the 166-D datasets used include 26,000,000 data points. The authors compared the performance of all the frameworks in terms of efficiency, cost, and availability. According to the study, the authors concluded that the loose coupling will be increasingly applied to research on electron cloud, and the parallel programming technology (MapReduce) framework may provide the user an interface with more convenient services and reduce unnecessary costs.

3 Big data generation and acquisition

We have introduced several key technologies related to big data, i.e., cloud computing, IoT, data center, and Hadoop. Next, we will focus on the value chain of big data, which can be generally divided into four phases: data generation, data acquisition, data storage, and data analysis. If we take data as a raw material, data generation and data acquisition are an exploitation process, data storage is a storage process, and data analysis is a production process that utilizes the raw material to create new value.

3.1 Data generation

Data generation is the first step of big data. Given Internet data as an example, huge amount of data in terms of searching entries, Internet forum posts, chatting records, and microblog messages, are generated. Those data are closely related to people's daily life, and have similar features of high value and low density. Such Internet data may be valueless individually, but, through the exploitation of accumulated big data, useful information such as habits and hobbies of users can be identified, and it is even possible to forecast users' behaviors and emotional moods.

Moreover, generated through longitudinal and/or distributed data sources, datasets are more large-scale, highly diverse, and complex. Such data sources include sensors, videos, clickstreams, and/or all other available data sources. At present, main sources of big data are the operation and trading information in enterprises, logistic and sensing information in the IoT, human interaction information and position information in the Internet world, and data generated in scientific research, etc. The information far surpasses the capacities of IT architectures and infrastructures of existing enterprises, while its real time requirement also greatly stresses the existing computing capacity.

3.1.1 Enterprise data

In 2013, IBM issued Analysis: *the Applications of Big Data to the Real World*, which indicates that the internal data of enterprises are the main sources of big data. The internal data of enterprises mainly consists of online trading data and online analysis data, most of which are historically static data and are managed by RDBMSs in a structured manner. In addition, production data, inventory data, sales data, and financial data, etc., also constitute enterprise internal data, which aims to capture informationized and data-driven activities in enterprises, so as to record all activities of enterprises in the form of internal data.

Over the past decades, IT and digital data have contributed a lot to improve the profitability of business departments. It is estimated that the business data volume of all companies in the world may double every 1.2 years [10], in which, the business turnover through the Internet, enterprises to enterprises, and enterprises to consumers per day will reach USD 450 billion [33]. The continuously increasing business data volume requires more effective real-time analysis so as to fully harvest its potential. For example, Amazon processes millions of terminal operations and more than 500,000 queries from third-party sellers per day [12]. Walmart processes one million customer trades per hour and such trading data are imported into a database with a capacity of over 2.5PB [3]. Akamai analyzes 75 million events per day for its target advertisements [13].

3.1.2 IoT data

As discussed, IoT is an important source of big data. Among smart cities constructed based on IoT, big data may come from industry, agriculture, traffic, transportation, medical care, public departments, and families, etc.

According to the processes of data acquisition and transmission in IoT, its network architecture may be divided into three layers: the sensing layer, the network layer, and the application layer. The sensing layer is responsible for data acquisition and mainly consists of sensor networks. The network layer is responsible for information transmission and processing, where close transmission may rely on sensor networks, and remote transmission shall depend on the Internet. Finally, the application layer support specific applications of IoT.

According to characteristics of Internet of Things, the data generated from IoT has the following features:

- *Large-scale data*: in IoT, masses of data acquisition equipments are distributedly deployed, which may acquire simple numeric data, e.g., location; or complex multimedia data, e.g., surveillance video. In order to meet the demands of analysis and processing, not only the currently acquired data, but also the historical data within a certain time frame should be stored. Therefore, data generated by IoT are characterized by large scales.
- *Heterogeneity*: because of the variety data acquisition devices, the acquired data is also different and such data features heterogeneity.
- *Strong time and space correlation*: in IoT, every data acquisition device are placed at a specific geographic location and every piece of data has time stamp. The time and space correlation are an important property of data from IoT. During data analysis and processing, time and space are also important dimensions for statistical analysis.
- *Effective data accounts for only a small portion of the big data*: a great quantity of noises may occur during the acquisition and transmission of data in IoT. Among datasets acquired by acquisition devices, only a small amount of abnormal data is valuable. For example, during the acquisition of traffic video, the few video frames that capture the violation of traffic regulations

and traffic accidents are more valuable than those only capturing the normal flow of traffic.

3.1.3 Bio-medical data

As a series of high-throughput bio-measurement technologies are innovatively developed in the beginning of the 21st century, the frontier research in the bio-medicine field also enters the era of big data. By constructing smart, efficient, and accurate analytical models and theoretical systems for bio-medicine applications, the essential governing mechanism behind complex biological phenomena may be revealed. Not only the future development of bio-medicine can be determined, but also the leading roles can be assumed in the development of a series of important strategic industries related to the national economy, people's livelihood, and national security, with important applications such as medical care, new drug R & D, and grain production (e.g., transgenic crops).

The completion of HGP (Human Genome Project) and the continued development of sequencing technology also lead to widespread applications of big data in the field. The masses of data generated by gene sequencing go through specialized analysis according to different application demands, to combine it with the clinical gene diagnosis and provide valuable information for early diagnosis and personalized treatment of disease. One sequencing of human gene may generate 100 600GB raw data. In the China National Genebank in Shenzhen, there are 1.3 million samples including 1.15 million human samples and 150,000 animal, plant, and microorganism samples. By the end of 2013, 10 million traceable biological samples will be stored, and by the end of 2015, this figure will reach 30 million. It is predictable that, with the development of bio-medicine technologies, gene sequencing will become faster and more convenient, and thus making big data of bio-medicine continuously grow beyond all doubt.

In addition, data generated from clinical medical care and medical R & D also rise quickly. For example, the University of Pittsburgh Medical Center (UPMC) has stored 2TB such data. Explorys, an American company, provides platforms to collocate clinical data, operation and maintenance data, and financial data. At present, about 13 million people's information have been collocated, with 44 articles of data at the scale of about 60TB, which will reach 70TB in 2013. Practice Fusion, another American company, manages electronic medical records of about 200,000 patients.

Apart from such small and medium-sized enterprises, other well-known IT companies, such as Google, Microsoft, and IBM have invested extensively in the research and computational analysis of methods related to high-throughput biological big data, for shares in the huge market as known as the "Next Internet." IBM forecasts, in the 2013 Strategy Conference, that with the sharp increase of medical images and electronic medical records, medical professionals may utilize big data to extract useful clinical information from masses of data to obtain a medical history and forecast treatment effects, thus improving patient care and reduce cost. It is anticipated that, by 2015, the average data volume of every hospital will increase from 167TB to 665TB.

3.1.4 Data generation from other fields

As scientific applications are increasing, the scale of datasets is gradually expanding, and the development of some disciplines greatly relies on the analysis of masses of data. Here, we examine several such applications. Although being in different scientific fields, the applications have similar and increasing demand on data analysis. The first example is related to computational biology. GenBank is a nucleotide sequence database maintained by the U.S. National Bio-Technology Innovation Center. Data in this database may double every 10 months. By August 2009, Genbank has more than 250 billion bases from 150,000 different organisms [34]. The second example is related to astronomy. Sloan Digital Sky Survey (SDSS), the biggest sky survey project in astronomy, has recorded 25TB data from 1998 to 2008. As the resolution of the telescope is improved, by 2004, the data volume generated per night will surpass 20TB. The last application is related to high-energy physics. In the beginning of 2008, the Atlas experiment of Large Hadron Collider (LHC) of European Organization for Nuclear Research generates raw data at 2PB/s and stores about 10TB processed data per year.

In addition, pervasive sensing and computing among nature, commercial, Internet, government, and social environments are generating heterogeneous data with unprecedented complexity. These datasets have their unique data characteristics in scale, time dimension, and data category. For example, mobile data were recorded with respect to positions, movement, approximation degrees, communications, multimedia, use of applications, and audio environment [108]. According to the application environment and requirements, such datasets into different categories, so as to select the proper and feasible solutions for big data.

3.2 Big data acquisition

As the second phase of the big data system, big data acquisition includes data collection, data transmission, and data pre-processing. During big data acquisition, once we collect the raw data, we shall utilize an efficient transmission mechanism to send it to a proper storage management system to support different analytical applications. The collected datasets may sometimes include much redundant or useless data, which unnecessarily increases storage space and affects the subsequent data analysis. For example, high redundancy is very common among datasets collected by sensors for environment monitoring. Data compression technology can be applied to reduce the redundancy. Therefore, data pre-processing operations are indispensable to ensure efficient data storage and exploitation.

3.2.1 Data collection

Data collection is to utilize special data collection techniques to acquire raw data from a specific data generation environment. Four common data collection methods are shown as follows.

- *Log files*: As one widely used data collection method, log files are record files automatically generated by the data source system, so as to record activities in designated file formats for subsequent analysis. Log files are typically used in nearly all digital devices. For example, web servers record in log files number of clicks, click rates, visits, and other property records of web users [35]. To capture activities of users at the web sites, web servers mainly include the following three log file formats: public log file format (NCSA), expanded log format (W3C), and IIS log format (Microsoft). All the three types of log files are in the ASCII text format. Databases other than text files may sometimes be used to store log information to improve the query efficiency of the massive log store [36, 37]. There are also some other log files based on data collection, including stock indicators in financial applications and determination of operating states in network monitoring and traffic management.
- *Sensing*: Sensors are common in daily life to measure physical quantities and transform physical quantities into readable digital signals for subsequent processing (and storage). Sensory data may be classified as sound wave, voice, vibration, automobile, chemical, current, weather, pressure, temperature, etc. Sensed information is transferred to a data collection point through wired or wireless networks. For applications that may be easily deployed and managed, e.g., video surveillance system [38], the wired sensor network is a convenient solution to acquire related information. Sometimes the accurate position of a specific phenomenon is unknown, and sometimes the monitored environment does not have the energy or communication infrastructures. Then wireless communication must be used to enable data transmission among sensor nodes under limited energy and communication capability. In recent years, WSNs have received considerable interest and have been applied to many applications, such as environmental research [39, 40], water quality monitoring [41], civil engineering [42, 43], and wildlife habit monitoring [44]. A WSN generally consists of a large number of geographically distributed sensor nodes, each being a micro device powered by battery. Such sensors are deployed at designated positions as required by the application to collect remote sensing data. Once the sensors are deployed, the base station will send control information for network configuration/management or data collection to sensor nodes. Based on such control information, the sensory data is assembled in different sensor nodes and sent back to the base station for further processing. Interested readers are referred to [45] for more detailed discussions.
- *Methods for acquiring network data*: At present, network data acquisition is accomplished using a combination of web crawler, word segmentation system, task system, and index system, etc. Web crawler is a program used by search engines for downloading and storing web pages [46]. Generally speaking, web crawler starts from the uniform resource locator (URL) of an initial web page to access other linked web pages, during which it stores and sequences all the retrieved URLs. Web crawler acquires a URL in the order of precedence through a URL queue and then downloads web pages, and identifies all URLs in the downloaded web pages, and extracts new URLs to be put in the queue. This process is repeated until the web crawler is stopped. Data acquisition through a web crawler is widely applied in applications based on web pages, such as search engines or web caching. Traditional web page extraction technologies feature multiple efficient solutions and considerable research has been done in this field. As more advanced web page applications are emerging, some extraction strategies are proposed in [47] to cope with rich Internet applications.

The current network data acquisition technologies mainly include traditional Libpcap-based packet capture technology, zero-copy packet capture technology, as well as some specialized network monitoring software such as Wireshark, SmartSniff, and WinNetCap.

- *Libpcap-based packet capture technology*: Libpcap (packet capture library) is a widely used network data packet capture function library. It is a general tool that does not depend on any specific system and is mainly used to capture data in the data link layer. It features simplicity, easy-to-use, and portability, but has a relatively low efficiency. Therefore, under a high-speed network environment, considerable packet losses may occur when Libpcap is used.

- *Zero-copy packet capture technology*: The so-called zero-copy (ZC) means that no copies between any internal memories occur during packet receiving and sending at a node. In sending, the data packets directly start from the user buffer of applications, pass through the network interfaces, and arrive at an external network. In receiving, the network interfaces directly send data packets to the user buffer. The basic idea of zero-copy is to reduce data copy times, reduce system calls, and reduce CPU load while ddatagrams are passed from network equipments to user program space. The zero-copy technology first utilizes direct memory access (DMA) technology to directly transmit network datagrams to an address space pre-allocated by the system kernel, so as to avoid the participation of CPU. In the meanwhile, it maps the internal memory of the datagrams in the system kernel to the that of the detection program, or builds a cache region in the user space and maps it to the kernel space. Then the detection program directly accesses the internal memory, so as to reduce internal memory copy from system kernel to user space and reduce the amount of system calls.
- *Mobile equipments*: At present, mobile devices are more widely used. As mobile device functions become increasingly stronger, they feature more complex and multiple means of data acquisition as well as more variety of data. Mobile devices may acquire geographical location information through positioning systems; acquire audio information through microphones; acquire pictures, videos, streetscapes, two-dimensional barcodes, and other multimedia information through cameras; acquire user gestures and other body language information through touch screens and gravity sensors. Over the years, wireless operators have improved the service level of the mobile Internet by acquiring and analyzing such information. For example, iPhone itself is a "mobile spy." It may collect wireless data and geographical location information, and then send such information back to Apple Inc. for processing, of which the user is not aware. Apart from Apple, smart phone operating systems such as Android of Google and Windows Phone of Microsoft can also collect information in the similar manner.

In addition to the aforementioned three data acquisition methods of main data sources, there are many other data collect methods or systems. For example, in scientific experiments, many special tools can be used to collect experimental data, such as magnetic spectrometers and radio telescopes. We may classify data collection methods from different perspectives. From the perspective of data sources, data collection methods can be classified into two categories: collection methods recording through data sources and collection methods recording through other auxiliary tools.

3.2.2 Data transportation

Upon the completion of raw data collection, data will be transferred to a data storage infrastructure for processing and analysis. As discussed in Section 2.3, big data is mainly stored in a data center. The data layout should be adjusted to improve computing efficiency or facilitate hardware maintenance. In other words, internal data transmission may occur in the data center. Therefore, data transmission consists of two phases: Inter-DCN transmissions and Intra-DCN transmissions.

- *Inter-DCN transmissions*: Inter-DCN transmissions are from data source to data center, which is generally achieved with the existing physical network infrastructure. Because of the rapid growth of traffic demands, the physical network infrastructure in most regions around the world are constituted by high-volumn, high-rate, and cost-effective optic fiber transmission systems. Over the past 20 years, advanced management equipment and technologies have been developed, such as IP-based wavelength division multiplexing (WDM) network architecture, to conduct smart control and management of optical fiber networks [48, 49]. WDM is a technology that multiplexes multiple optical carrier signals with different wave lengths and couples them to the same optical fiber of the optical link. In such technology, lasers with different wave lengths carry different signals. By far, the backbone network have been deployed with WDM optical transmission systems with single channel rate of 40Gb/s. At present, 100Gb/s commercial interface are available and 100Gb/s systems (or TB/s systems) will be available in the near future [50]. However, traditional optical transmission technologies are limited by the bandwidth of the electronic bottleneck [51]. Recently, orthogonal frequency-division multiplexing (OFDM), initially designed for wireless systems, is regarded as one of the main candidate technologies for future high-speed optical transmission. OFDM is a multi-carrier parallel transmission technology. It segments a high-speed data flow to transform it into low-speed sub-data-flows to be transmitted over multiple orthogonal sub-carriers [52]. Compared with fixed channel spacing of WDM, OFDM allows sub-channel frequency spectrums to overlap with each other [53]. Therefore, it is a flexible and efficient optical networking technology.
- *Intra-DCN Transmissions*: Intra-DCN transmissions are the data communication flows within data centers. Intra-DCN transmissions depend on the communication

mechanism within the data center (i.e., on physical connection plates, chips, internal memories of data servers, network architectures of data centers, and communication protocols). A data center consists of multiple integrated server racks interconnected with its internal connection networks. Nowadays, the internal connection networks of most data centers are fat-tree, two-layer or three-layer structures based on multi-commodity network flows [51, 54]. In the two-layer topological structure, the racks are connected by 1Gbps top rack switches (TOR) and then such top rack switches are connected with 10Gbps aggregation switches in the topological structure. The three-layer topological structure is a structure augmented with one layer on the top of the two-layer topological structure and such layer is constituted by 10Gbps or 100Gbps core switches to connect aggregation switches in the topological structure. There are also other topological structures which aim to improve the data center networks [55–58]. Because of the inadequacy of electronic packet switches, it is difficult to increase communication bandwidths while keeps energy consumption is low. Over the years, due to the huge success achieved by optical technologies, the optical interconnection among the networks in data centers has drawn great interest. Optical interconnection is a high-throughput, low-delay, and low-energy-consumption solution. At present, optical technologies are only used for point-to-point links in data centers. Such optical links provide connection for the switches using the low-cost multi-mode fiber (MMF) with 10Gbps data rate. Optical interconnection (switching in the optical domain) of networks in data centers is a feasible solution, which can provide Tbps-level transmission bandwidth with low energy consumption. Recently, many optical interconnection plans are proposed for data center networks [59]. Some plans add optical paths to upgrade the existing networks, and other plans completely replace the current switches [59–64]. As a strengthening technology, Zhou et al. in [65] adopt wireless links in the 60GHz frequency band to strengthen wired links. Network virtualization should also be considered to improve the efficiency and utilization of data center networks.

3.2.3 Data pre-processing

Because of the wide variety of data sources, the collected datasets vary with respect to noise, redundancy, and consistency, etc., and it is undoubtedly a waste to store meaningless data. In addition, some analytical methods have serious requirements on data quality. Therefore, in order to enable effective data analysis, we shall pre-process data under many circumstances to integrate the data from different sources, which can not only reduces storage expense, but also improves analysis accuracy. Some relational data pre-processing techniques are discussed as follows.

- *Integration*: data integration is the cornerstone of modern commercial informatics, which involves the combination of data from different sources and provides users with a uniform view of data [66]. This is a mature research field for traditional database. Historically, two methods have been widely recognized: data warehouse and data federation. Data warehousing includes a process named ETL (Extract, Transform and Load). Extraction involves connecting source systems, selecting, collecting, analyzing, and processing necessary data. Transformation is the execution of a series of rules to transform the extracted data into standard formats. Loading means importing extracted and transformed data into the target storage infrastructure. Loading is the most complex procedure among the three, which includes operations such as transformation, copy, clearing, standardization, screening, and data organization. A virtual database can be built to query and aggregate data from different data sources, but such database does not contain data. On the contrary, it includes information or metadata related to actual data and its positions. Such two "storage-reading" approaches do not satisfy the high performance requirements of data flows or search programs and applications. Compared with queries, data in such two approaches is more dynamic and must be processed during data transmission. Generally, data integration methods are accompanied with flow processing engines and search engines [30, 67].
- *Cleaning*: data cleaning is a process to identify inaccurate, incomplete, or unreasonable data, and then modify or delete such data to improve data quality. Generally, data cleaning includes five complementary procedures [68]: defining and determining error types, searching and identifying errors, correcting errors, documenting error examples and error types, and modifying data entry procedures to reduce future errors. During cleaning, data formats, completeness, rationality, and restriction shall be inspected. Data cleaning is of vital importance to keep the data consistency, which is widely applied in many fields, such as banking, insurance, retail industry, telecommunications, and traffic control.

  In e-commerce, most data is electronically collected, which may have serious data quality problems. Classic data quality problems mainly come from software defects, customized errors, or system misconfiguration. Authors in [69] discussed data cleaning in e-commerce by crawlers and regularly re-copying customer and account information.

In [70], the problem of cleaning RFID data was examined. RFID is widely used in many applications, e.g., inventory management and target tracking. However, the original RFID features low quality, which includes a lot of abnormal data limited by the physical design and affected by environmental noises. In [71], a probability model was developed to cope with data loss in mobile environments. Khoussainova et al. in [72] proposed a system to automatically correct errors of input data by defining global integrity constraints.

Herbert et al. [73] proposed a framework called BIO-AJAX to standardize biological data so as to conduct further computation and improve search quality. With BIO-AJAX, some errors and repetitions may be eliminated, and common data mining technologies can be executed more effectively.

– *Redundancy elimination*: data redundancy refers to data repetitions or surplus, which usually occurs in many datasets. Data redundancy can increase the unnecessary data transmission expense and cause defects on storage systems, e.g., waste of storage space, leading to data inconsistency, reduction of data reliability, and data damage. Therefore, various redundancy reduction methods have been proposed, such as redundancy detection, data filtering, and data compression. Such methods may apply to different datasets or application environments. However, redundancy reduction may also bring about certain negative effects. For example, data compression and decompression cause additional computational burden. Therefore, the benefits of redundancy reduction and the cost should be carefully balanced. Data collected from different fields will increasingly appear in image or video formats. It is well-known that images and videos contain considerable redundancy, including temporal redundancy, spacial redundancy, statistical redundancy, and sensing redundancy. Video compression is widely used to reduce redundancy in video data, as specified in the many video coding standards (MPEG-2, MPEG-4, H.263, and H.264/AVC). In [74], the authors investigated the problem of video compression in a video surveillance system with a video sensor network. The authors propose a new MPEG-4 based method by investigating the contextual redundancy related to background and foreground in a scene. The low complexity and the low compression ratio of the proposed approach were demonstrated by the evaluation results.

On generalized data transmission or storage, repeated data deletion is a special data compression technology, which aims to eliminate repeated data copies [75]. With repeated data deletion, individual data blocks or data segments will be assigned with identifiers (e.g., using a hash algorithm) and stored, with the identifiers added to the identification list. As the analysis of repeated data deletion continues, if a new data block has an identifier that is identical to that listed in the identification list, the new data block will be deemed as redundant and will be replaced by the corresponding stored data block. Repeated data deletion can greatly reduce storage requirement, which is particularly important to a big data storage system. Apart from the aforementioned data pre-processing methods, specific data objects shall go through some other operations such as feature extraction. Such operation plays an important role in multimedia search and DNA analysis [76–78]. Usually high-dimensional feature vectors (or high-dimensional feature points) are used to describe such data objects and the system stores the dimensional feature vectors for future retrieval. Data transfer is usually used to process distributed heterogeneous data sources, especially business datasets [79]. As a matter of fact, in consideration of various datasets, it is non-trivial, or impossible, to build a uniform data pre-processing procedure and technology that is applicable to all types of datasets. on the specific feature, problem, performance requirements, and other factors of the datasets should be considered, so as to select a proper data pre-processing strategy.

4 Big data storage

The explosive growth of data has more strict requirements on storage and management. In this section, we focus on the storage of big data. Big data storage refers to the storage and management of large-scale datasets while achieving reliability and availability of data accessing. We will review important issues including massive storage systems, distributed storage systems, and big data storage mechanisms. On one hand, the storage infrastructure needs to provide information storage service with reliable storage space; on the other hand, it must provide a powerful access interface for query and analysis of a large amount of data.

Traditionally, as auxiliary equipment of server, data storage device is used to store, manage, look up, and analyze data with structured RDBMSs. With the sharp growth of data, data storage device is becoming increasingly more important, and many Internet companies pursue big capacity of storage to be competitive. Therefore, there is a compelling need for research on data storage.

4.1 Storage system for massive data

Various storage systems emerge to meet the demands of massive data. Existing massive storage technologies can be classified as Direct Attached Storage (DAS) and network storage, while network storage can be further classified into Network Attached Storage (NAS) and Storage Area Network (SAN).

In DAS, various harddisks are directly connected with servers, and data management is server-centric, such that storage devices are peripheral equipments, each of which takes a certain amount of I/O resource and is managed by an individual application software. For this reason, DAS is only suitable to interconnect servers with a small scale. However, due to its low scalability, DAS will exhibit undesirable efficiency when the storage capacity is increased, i.e., the upgradeability and expandability are greatly limited. Thus, DAS is mainly used in personal computers and small-sized servers.

Network storage is to utilize network to provide users with a union interface for data access and sharing. Network storage equipment includes special data exchange equipments, disk array, tap library, and other storage media, as well as special storage software. It is characterized with strong expandability.

NAS is actually an auxillary storage equipment of a network. It is directly connected to a network through a hub or switch through TCP/IP protocols. In NAS, data is transmitted in the form of files. Compared to DAS, the I/O burden at a NAS server is reduced extensively since the server accesses a storage device indirectly through a network.

While NAS is network-oriented, SAN is especially designed for data storage with a scalable and bandwidth intensive network, e.g., a high-speed network with optical fiber connections. In SAN, data storage management is relatively independent within a storage local area network, where multipath based data switching among any internal nodes is utilized to achieve a maximum degree of data sharing and data management.

From the organization of a data storage system, DAS, NAS, and SAN can all be divided into three parts: (i) disc array: it is the foundation of a storage system and the fundamental guarantee for data storage; (ii) connection and network sub-systems, which provide connection among one or more disc arrays and servers; (iii) storage management software, which handles data sharing, disaster recovery, and other storage management tasks of multiple servers.

4.2 Distributed storage system

The first challenge brought about by big data is how to develop a large scale distributed storage system for efficiently data processing and analysis. To use a distributed system to store massive data, the following factors should be taken into consideration:

- *Consistency*: a distributed storage system requires multiple servers to cooperatively store data. As there are more servers, the probability of server failures will be larger. Usually data is divided into multiple pieces to be stored at different servers to ensure availability in case of server failure. However, server failures and parallel storage may cause inconsistency among different copies of the same data. Consistency refers to assuring that multiple copies of the same data are identical.
- *Availability*: a distributed storage system operates in multiple sets of servers. As more servers are used, server failures are inevitable. It would be desirable if the entire system is not seriously affected to satisfy customer's requests in terms of reading and writing. This property is called availability.
- *Partition Tolerance*: multiple servers in a distributed storage system are connected by a network. The network could have link/node failures or temporary congestion. The distributed system should have a certain level of tolerance to problems caused by network failures. It would be desirable that the distributed storage still works well when the network is partitioned.

Eric Brewer proposed a CAP [80, 81] theory in 2000, which indicated that a distributed system could not simultaneously meet the requirements on consistency, availability, and partition tolerance; at most two of the three requirements can be satisfied simultaneously. Seth Gilbert and Nancy Lynch from MIT proved the correctness of CAP theory in 2002. Since consistency, availability, and partition tolerance could not be achieved simultaneously, we can have a CA system by ignoring partition tolerance, a CP system by ignoring availability, and an AP system that ignores consistency, according to different design goals. The three systems are discussed in the following.

CA systems do not have partition tolerance, i.e, they could not handle network failures. Therefore, CA systems are generally deemed as storage systems with a single server, such as the traditional small-scale relational databases. Such systems feature single copy of data, such that consistency is easily ensured. Availability is guaranteed by the excellent design of relational databases. However, since CA systems could not handle network failures, they could not be expanded to use many servers. Therefore, most large-scale storage systems are CP systems and AP systems.

Compared with CA systems, CP systems ensure partition tolerance. Therefore, CP systems can be expanded to become distributed systems. CP systems generally maintain several copies of the same data in order to ensure a level of fault tolerance. CP systems also ensure data consistency, i.e., multiple copies of the same data are guaranteed to be completely identical. However, CP could not ensure sound availability because of the high cost for consistency assurance. Therefore, CP systems are useful for the scenarios with moderate load but stringent requirements on data accuracy (e.g., trading data). BigTable and Hbase are two popular CP systems.

AP systems also ensure partition tolerance. However, AP systems are different from CP systems in that AP systems also ensure availability. However, AP systems only ensure eventual consistency rather than strong consistency in the previous two systems. Therefore, AP systems only apply to the scenarios with frequent requests but not very high requirements on accuracy. For example, in online Social Networking Services (SNS) systems, there are many concurrent visits to the data but a certain amount of data errors are tolerable. Furthermore, because AP systems ensure eventual consistency, accurate data can still be obtained after a certain amount of delay. Therefore, AP systems may also be used under the circumstances with no stringent realtime requirements. Dynamo and Cassandra are two popular AP systems.

4.3 Storage mechanism for big data

Considerable research on big data promotes the development of storage mechanisms for big data. Existing storage mechanisms of big data may be classified into three bottom-up levels: (i) file systems, (ii) databases, and (iii) programming models.

File systems are the foundation of the applications at upper levels. Google's GFS is an expandable distributed file system to support large-scale, distributed, data-intensive applications [25]. GFS uses cheap commodity servers to achieve fault-tolerance and provides customers with high-performance services. GFS supports large-scale file applications with more frequent reading than writing. However, GFS also has some limitations, such as a single point of failure and poor performances for small files. Such limitations have been overcome by Colossus [82], the successor of GFS.

In addition, other companies and researchers also have their solutions to meet the different demands for storage of big data. For example, HDFS and Kosmosfs are derivatives of open source codes of GFS. Microsoft developed Cosmos [83] to support its search and advertisement business. Facebook utilizes Haystack [84] to store the large amount of small-sized photos. Taobao also developed TFS and FastDFS. In conclusion, distributed file systems have been relatively mature after years of development and business operation. Therefore, we will focus on the other two levels in the rest of this section.

4.3.1 Database technology

The database technology has been evolving for more than 30 years. Various database systems are developed to handle datasets at different scales and support various applications. Traditional relational databases cannot meet the challenges on categories and scales brought about by big data. NoSQL databases (i.e., non traditional relational databases) are becoming more popular for big data storage. NoSQL databases feature flexible modes, support for simple and easy copy, simple API, eventual consistency, and support of large volume data. NoSQL databases are becoming the core technology for of big data. We will examine the following three main NoSQL databases in this section: Key-value databases, column-oriented databases, and document-oriented databases, each based on certain data models.

- *Key-value Databases*: Key-value Databases are constituted by a simple data model and data is stored corresponding to key-values. Every key is unique and customers may input queried values according to the keys. Such databases feature a simple structure and the modern key-value databases are characterized with high expandability and shorter query response time than those of relational databases. Over the past few years, many key-value databases have appeared as motivated by Amazon's Dynamo system [85]. We will introduce Dynamo and several other representative key-value databases.
    - *Dynamo*: Dynamo is a highly available and expandable distributed key-value data storage system . It is used to store and manage the status of some core services, which can be realized with key access, in the Amazon e-Commerce Platform. The public mode of relational databases may generate invalid data and limit data scale and availability, while Dynamo can resolve these problems with a simple key-object interface, which is constituted by simple reading and writing operation. Dynamo achieves elasticity and availability through the data partition, data copy, and object edition mechanisms. Dynamo partition plan relies on Consistent Hashing [86], which has a main advantage that node passing only affects directly adjacent nodes and do not affect other nodes, to divide the load for multiple main storage machines. Dynamo copies data to N sets of servers, in which N is a configurable parameter in order to achieve

high availability and durability. Dynamo system also provides eventual consistency, so as to conduct asynchronous update on all copies.

  - *Voldemort*: Voldemort is also a key-value storage system, which was initially developed for and is still used by LinkedIn. Key words and values in Voldemort are composite objects constituted by tables and images. Voldemort interface includes three simple operations: reading, writing, and deletion, all of which are confirmed by key words. Voldemort provides asynchronous updating concurrent control of multiple editions but does not ensure data consistency. However, Voldemort supports optimistic locking for consistent multi-record updating. When conflict happens between the updating and any other operations, the updating operation will quit. The data copy mechanism of Voldmort is the same as that of Dynamo. Voldemort not only stores data in RAM but allows data be inserted into a storage engine. Especially, Voldemort supports two storage engines including Berkeley DB and Random Access Files.

The key-value database emerged a few years ago. Deeply influenced by Amazon Dynamo DB, other key-value storage systems include Redis, Tokyo Canbinet and Tokyo Tyrant, Memcached and Memcache DB, Riak and Scalaris, all of which provide expandability by distributing key words into nodes. Voldemort, Riak, Tokyo Cabinet, and Memecached can utilize attached storage devices to store data in RAM or disks. Other storage systems store data at RAM and provide disk backup, or rely on copy and recovery to avoid backup.

- *Column-oriented Database*: The column-oriented databases store and process data according to columns other than rows. Both columns and rows are segmented in multiple nodes to realize expandability. The column-oriented databases are mainly inspired by Google's BigTable. In this Section, we first discuss BigTable and then introduce several derivative tools.

  - *BigTable*: BigTable is a distributed, structured data storage system, which is designed to process the large-scale (PB class) data among thousands commercial servers [87]. The basic data structure of Bigtable is a multi-dimension sequenced mapping with sparse, distributed, and persistent storage. Indexes of mapping are row key, column key, and timestamps, and every value in mapping is an unanalyzed byte array. Each row key in BigTable is a 64KB character string. By lexicographical order, rows are stored and continually segmented into Tablets (i.e., units of distribution) for load balance. Thus, reading a short row of data can be highly effective, since it only involves communication with a small portion of machines. The columns are grouped according to the prefixes of keys, and thus forming column families. These column families are the basic units for access control. The timestamps are 64-bit integers to distinguish different editions of cell values. Clients may flexibly determine the number of cell editions stored. These editions are sequenced in the descending order of timestamps, so the latest edition will always be read.

    The BigTable API features the creation and deletion of Tablets and column families as well as modification of metadata of clusters, tables, and column families. Client applications may insert or delete values of BigTable, query values from columns, or browse sub-datasets in a table. Bigtable also supports some other characteristics, such as transaction processing in a single row. Users may utilize such features to conduct more complex data processing.

    Every procedure executed by BigTable includes three main components: Master server, Tablet server, and client library. Bigtable only allows one set of Master server be distributed to be responsible for distributing tablets for Tablet server, detecting added or removed Tablet servers, and conducting load balance. In addition, it can also modify BigTable schema, e.g., creating tables and column families, and collecting garbage saved in GFS as well as deleted or disabled files, and using them in specific BigTable instances. Every tablet server manages a Tablet set and is responsible for the reading and writing of a loaded Tablet. When Tablets are too big, they will be segmented by the server. The application client library is used to communicate with BigTable instances.

    BigTable is based on many fundamental components of Google, including GFS [25], cluster management system, SSTable file format, and Chubby [88]. GFS is use to store data and log files. The cluster management system is responsible for task scheduling, resources sharing, processing of machine failures, and monitoring of machine statuses. SSTable file format is used to store BigTable data internally, and it provides mapping between persistent, sequenced, and unchangeable keys and values as any byte strings. BigTable utilizes Chubby for the following tasks in server: 1) ensure there is at most one active Master copy at any time; 2) store the bootstrap location of BigTable data; 3) look up Tablet server; 4) conduct error recovery in case of Table server failures; 5) store BigTable schema information; 6) store the access control table.

- *Cassandra*: Cassandra is a distributed storage system to manage the huge amount of structured data distributed among multiple commercial servers [89]. The system was developed by Facebook and became an open source tool in 2008. It adopts the ideas and concepts of both Amazon Dynamo and Google BigTable, especially integrating the distributed system technology of Dynamo with the BigTable data model. Tables in Cassandra are in the form of distributed four-dimensional structured mapping, where the four dimensions including row, column, column family, and super column. A row is distinguished by a string-key with arbitrary length. No matter the amount of columns to be read or written, the operation on rows is an auto. Columns may constitute clusters, which is called column families, and are similar to the data model of Bigtable. Cassandra provides two kinds of column families: column families and super columns. The super column includes arbitrary number of columns related to same names. A column family includes columns and super columns, which may be continuously inserted to the column family during runtime. The partition and copy mechanisms of Cassandra are very similar to those of Dynamo, so as to achieve consistency.
- *Derivative tools of BigTable*: since the BigTable code cannot be obtained through the open source license, some open source projects compete to implement the BigTable concept to develop similar systems, such as HBase and Hypertable.

  HBase is a BigTable cloned version programmed with Java and is a part of Hadoop of Apache's MapReduce framework [90]. HBase replaces GFS with HDFS. It writes updated contents into RAM and regularly writes them into files on disks. The row operations are atomic operations, equipped with row-level locking and transaction processing, which is optional for large scale. Partition and distribution are transparently operated and have space for client hash or fixed key.

  HyperTable was developed similar to BigTable to obtain a set of high-performance, expandable, distributed storage and processing systems for structured and unstructured data [91]. HyperTable relies on distributed file systems, e.g. HDFS and distributed lock manager. Data representation, processing, and partition mechanism are similar to that in BigTable. HyperTable has its own query language, called HyperTable query language (HQL), and allows users to create, modify, and query underlying tables.

Since the column-oriented storage databases mainly emulate BigTable, their designs are all similar, except for the concurrency mechanism and several other features. For example, Cassandra emphasizes weak consistency of concurrent control of multiple editions while HBase and HyperTable focus on strong consistency through locks or log records.

- *Document Database*: Compared with key-value storage, document storage can support more complex data forms. Since documents do not follow strict modes, there is no need to conduct mode migration. In addition, key-value pairs can still be saved. We will examine three important representatives of document storage systems, i.e., MongoDB, SimpleDB, and CouchDB.

  - *MongoDB*: MongoDB is open-source and document-oriented database [92]. MongoDB stores documents as Binary JSON (BSON) objects [93], which is similar to object. Every document has an ID field as the primary key. Query in MongoDB is expressed with syntax similar to JSON. A database driver sends the query as a BSON object to MongoDB. The system allows query on all documents, including embedded objects and arrays. To enable rapid query, indexes can be created in the queryable fields of documents. The copy operation in MongoDB can be executed with log files in the main nodes that support all the high-level operations conducted in the database. During copying, the slavers query all the writing operations since the last synchronization to the master and execute operations in log files in local databases. MongoDB supports horizontal expansion with automatic sharing to distribute data among thousands of nodes by automatically balancing load and failover.

- *SimpleDB*: SimpleDB is a distributed database and is a web service of Amazon [94]. Data in SimpleDB is organized into various domains in which data may be stored, acquired, and queried. Domains include different properties and name/value pair sets of projects. Date is copied to different machines at different data centers in order to ensure data safety and improve performance. This system does not support automatic partition and thus could not be expanded with the change of data volume. SimpleDB allows users to query with SQL. It is worth noting that SimpleDB can assure eventual consistency but does not support to Muti-Version Concurrency Control (MVCC). Therefore, conflicts therein could not be detected from the client side.
- *CouchDB*: Apache CouchDB is a document-oriented database written in Erlang [95]. Data in CouchDB is organized into documents consisting of fields named by keys/names and values, which are stored and accessed as JSON objects. Every document is provided with a unique identifier. CouchDB allows access to database documents through the RESTful HTTP API. If a document needs to be modified, the client must download the entire document to modify it, and then send it back to the database. After a document is rewritten once, the identifier will be updated. CouchDB utilizes the optimal copying to obtain scalability without a sharing mechanism. Since various CouchDBs may be executed along with other transactions simultaneously, any kinds of Replication Topology can be built. The consistency of CouchDB relies on the copying mechanism. CouchDB supports MVCC with historical Hash records.

Big data are generally stored in hundreds and even thousands of commercial servers. Thus, the traditional parallel models, such as Message Passing Interface (MPI) and Open Multi-Processing (OpenMP), may not be adequate to support such large-scale parallel programs. Recently, some proposed parallel programming models effectively improve the performance of NoSQL and reduce the performance gap to relational databases. Therefore, these models have become the cornerstone for the analysis of massive data.

- *MapReduce*: MapReduce [22] is a simple but powerful programming model for large-scale computing using a large number of clusters of commercial PCs to achieve automatic parallel processing and distribution. In MapReduce, computing model only has two functions, i.e., Map and Reduce, both of which are programmed by users. The Map function processes input key-value pairs and generates intermediate key-value pairs. Then, MapReduce will combine all the intermediate values related to the same key and transmit them to the Reduce function, which further compress the value set into a smaller set. MapReduce has the advantage that it avoids the complicated steps for developing parallel applications, e.g., data scheduling, fault-tolerance, and inter-node communications. The user only needs to program the two functions to develop a parallel application. The initial MapReduce framework did not support multiple datasets in a task, which has been mitigated by some recent enhancements [96, 97].

  Over the past decades, programmers are familiar with the advanced declarative language of SQL, often used in a relational database, for task description and dataset analysis. However, the succinct MapReduce framework only provides two nontransparent functions, which cannot cover all the common operations. Therefore, programmers have to spend time on programming the basic functions, which are typically hard to be maintained and reused. In order to improve the programming efficiency, some advanced language systems have been proposed, e.g., Sawzall [98] of Google, Pig Latin [99] of Yahoo, Hive [100] of Facebook, and Scope [87] of Microsoft.
- *Dryad*: Dryad [101] is a general-purpose distributed execution engine for processing parallel applications of coarse-grained data. The operational structure of Dryad is a directed acyclic graph, in which vertexes represent programs and edges represent data channels. Dryad executes operations on the vertexes in clusters and transmits data via data channels, including documents, TCP connections, and shared-memory FIFO. During operation, resources in a logic operation graph are automatically map to physical resources.

  The operation structure of Dryad is coordinated by a central program called job manager, which can be executed in clusters or workstations through network. A job manager consists of two parts: 1) application codes which are used to build a job communication graph, and 2) program library codes that are used to arrange available resources. All kinds of data are directly transmitted between vertexes. Therefore, the job manager is only responsible for decision-making, which does not obstruct any data transmission.

  In Dryad, application developers can flexibly choose any directed acyclic graph to describe the communication modes of the application and express data transmission mechanisms. In addition, Dryad allows vertexes to use any amount of input and output data, while MapReduce supports only one input and output set.

DryadLINQ [102] is the advanced language of Dryad and is used to integrate the aforementioned SQL-like language execution environment.

- *All-Pairs*: All-Pairs [103] is a system specially designed for biometrics, bio-informatics, and data mining applications. It focuses on comparing element pairs in two datasets by a given function. All-Pairs can be expressed as three-tuples (Set A, Set B, and Function F), in which Function F is utilized to compare all elements in Set A and Set B. The comparison result is an output matrix **M**, which is also called the Cartesian product or cross join of Set A and Set B.

  All-Pairs is implemented in four phases: system modeling, distribution of input data, batch job management, and result collection. In Phase I, an approximation model of system performance will be built to evaluate how much CPU resource is needed and how to conduct job partition. In Phase II, a spanning tree is built for data transmissions, which makes the workload of every partition retrieve input data effectively. In Phase III, after the data flow is delivered to proper nodes, the All-Pairs engine will build a batch-processing submission for jobs in partitions, while sequencing them in the batch processing system, and formulating a node running command to acquire data. In the last phase, after the job completion of the batch processing system, the extraction engine will collect results and combine them in a proper structure, which is generally a single file list, in which all results are put in order.
- *Pregel*: The Pregel [104] system of Google facilitates the processing of large-sized graphs, e.g., analysis of network graphs and social networking services. A computational task is expressed by a directed graph constituted by vertexes and directed edges. Every vertex is related to a modifiable and user-defined value, and every directed edge related to a source vertex is constituted by the user-defined value and the identifier of a target vertex. When the graph is built, the program conducts iterative calculations, which is called supersteps among which global synchronization points are set until algorithm completion and output completion. In every superstep, vertex computations are parallel, and every vertex executes the same user-defined function to express a given algorithm logic. Every vertex may modify its and its output edges status, receive a message sent from the previous superstep, send the message to other vertexes, and even modify the topological structure of the entire graph. Edges are not provided with corresponding computations. Functions of every vertex may be removed by suspension. When all vertexes are in an inactive status without any message to transmit, the entire program execution is completed. The Pregel program output is a set consisting of the values output from all the vertexes. Generally speaking, the input and output of Pregel program are isomorphic directed graphs.

Inspired by the above programming models, other researches have also focused on programming modes for more complex computational tasks, e.g., iterative computations [105, 106], fault-tolerant memory computations [107], incremental computations [108], and flow control decision-making related to data [109].

5 Big data analysis

The analysis of big data mainly involves analytical methods for traditional data and big data, analytical architecture for big data, and software used for mining and analysis of big data. Data analysis is the final and the most important phase in the value chain of big data, with the purpose of extracting useful values, providing suggestions or decisions. Different levels of potential values can be generated through the analysis of datasets in different fields [10]. However, data analysis is a broad area, which frequently changes and is extremely complex. In this section, we introduce the methods, architectures and tools for big data analysis.

5.1 Traditional data analysis

Traditional data analysis means to use proper statistical methods to analyze massive data, to concentrate, extract, and refine useful data hidden in a batch of chaotic datasets, and to identify the inherent law of the subject matter, so as to maximize the value of data. Data analysis plays a huge guidance role in making development plans for a country, understanding customer demands for commerce, and predicting market trend for enterprises. Big data analysis can be deemed as the analysis technique for a special kind of data. Therefore, many traditional data analysis methods may still be utilized for big data analysis. Several representative traditional data analysis methods are examined in the following, many of which are from statistics and computer science.

- *Cluster Analysis*: is a statistical method for grouping objects, and specifically, classifying objects according to some features. Cluster analysis is used to differentiate objects with particular features and divide them into some categories (clusters) according to these features, such that objects in the same category will have high homogeneity while different categories will have high heterogeneity. Cluster analysis is an unsupervised study method without training data.
- *Factor Analysis*: is basically targeted at describing the relation among many elements with only a few factors, i.e., grouping several closely related variables into a factor, and the few factors are then used to reveal the most information of the original data.

- *Correlation Analysis*: is an analytical method for determining the law of relations, such as correlation, correlative dependence, and mutual restriction, among observed phenomena and accordingly conducting forecast and control. Such relations may be classified into two types: (i) function, reflecting the strict dependence relationship among phenomena, which is also called a definitive dependence relationship; (ii) correlation, some undetermined or inexact dependence relations, and the numerical value of a variable may correspond to several numerical values of the other variable, and such numerical values present a regular fluctuation surrounding their mean values.
- *Regression Analysis*: is a mathematical tool for revealing correlations between one variable and several other variables. Based on a group of experiments or observed data, regression analysis identifies dependence relationships among variables hidden by randomness. Regression analysis may make complex and undetermined correlations among variables to be simple and regular.
- *A/B Testing*: also called bucket testing. It is a technology for determining how to improve target variables by comparing the tested group. Big data will require a large number of tests to be executed and analyzed.
- *Statistical Analysis*: Statistical analysis is based on the statistical theory, a branch of applied mathematics. In statistical theory, randomness and uncertainty are modeled with Probability Theory. Statistical analysis can provide a description and an inference for big data. Descriptive statistical analysis can summarize and describe datasets, while inferential statistical analysis can draw conclusions from data subject to random variations. Statistical analysis is widely applied in the economic and medical care fields [110].
- *Data Mining Algorithms*: Data mining is a process for extracting hidden, unknown, but potentially useful information and knowledge from massive, incomplete, noisy, fuzzy, and random data. In 2006, The IEEE International Conference on Data Mining Series (ICDM) identified ten most influential data mining algorithms through a strict selection procedure [111], including C4.5, k-means, SVM, Apriori, EM, Naive Bayes, and Cart, etc. These ten algorithms cover classification, clustering, regression, statistical learning, association analysis, and linking mining, all of which are the most important problems in data mining research.

5.2 Big data analytic methods

In the dawn of the big data era, people are concerned how to rapidly extract key information from massive data so as to bring values for enterprises and individuals. At present, the main processing methods of big data are shown as follows.

- *Bloom Filter*: Bloom Filter consists of a series of Hash functions. The principle of Bloom Filter is to store Hash values of data other than data itself by utilizing a bit array, which is in essence a bitmap index that uses Hash functions to conduct lossy compression storage of data. It has such advantages as high space efficiency and high query speed, but also has some disadvantages in misrecognition and deletion.
- *Hashing*: it is a method that essentially transforms data into shorter fixed-length numerical values or index values. Hashing has such advantages as rapid reading, writing, and high query speed, but it is hard to find a sound Hash function.
- *Index*: index is always an effective method to reduce the expense of disk reading and writing, and improve insertion, deletion, modification, and query speeds in both traditional relational databases that manage structured data, and other technologies that manage semi-structured and unstructured data. However, index has a disadvantage that it has the additional cost for storing index files which should be maintained dynamically when data is updated.
- *Triel*: also called trie tree, a variant of Hash Tree. It is mainly applied to rapid retrieval and word frequency statistics. The main idea of Triel is to utilize common prefixes of character strings to reduce comparison on character strings to the greatest extent, so as to improve query efficiency.
- *Parallel Computing*: compared to traditional serial computing, parallel computing refers to simultaneously utilizing several computing resources to complete a computation task. Its basic idea is to decompose a problem and assign them to several separate processes to be independently completed, so as to achieve co-processing. Presently, some classic parallel computing models include MPI (Message Passing Interface), MapReduce, and Dryad (See a comparison in Table 1).

Although the parallel computing systems or tools, such as MapReduce or Dryad, are useful for big data analysis, they are low levels tools that are hard to learn and use. Therefore, some high-level parallel programming tools or languages are being developed based on these systems. Such high-level languages include Sawzall, Pig, and Hive used for MapReduce, as well as Scope and DryadLINQ used for Dryad.

5.3 Architecture for big data analysis

Because of the 4Vs of big data, different analytical architectures shall be considered for different application requirements.

Table 1 Comparison of MPI, MapReduce and Dryad

	MPI	MapReduce	Dryad
Deployment	Computing node and data storage arranged separately (Data should be moved computing node)	Computing and data storage arranged at the same node (Computing should be close to data)	Computing and data storage arranged at the same node (Computing should be close to data)

Resource management/ scheduling	–	Workqueue(google)
		HOD(Yahoo)
			Not clear

Low level programming	MPI API	MapReduce API	Dryad API
High level programming	–	Pig, Hive, Jaql, · · ·	Scope, DryadLINQ
Data storage	The local file system, NFS, · · ·	GFS(google) HDFS(Hadoop), KFS Amazon S3, · · ·	NTFS, Cosmos DFS
Task partitioning	User manually partition the tasks	Automation	Automation
Communication	Messaging, Remote memory access	Files(Local FS, DFS)	Files, TCP Pipes, Shared-memory FIFOs
Fault-tolerant	Checkpoint	Task re-execute	Task re-execute

5.3.1 Real-time vs. offline analysis

According to timeliness requirements, big data analysis can be classified into real-time analysis and off-line analysis.

– *Real-time analysis*: is mainly used in E-commerce and finance. Since data constantly changes, rapid data analysis is needed and analytical results shall be returned with a very short delay. The main existing architectures of real-time analysis include (i) parallel processing clusters using traditional relational databases, and (ii) memory-based computing platforms. For example, Greenplum from EMC and HANA from SAP are both real-time analysis architectures.

– *Offline analysis*: is usually used for applications without high requirements on response time, e.g., machine learning, statistical analysis, and recommendation algorithms. Offline analysis generally conducts analysis by importing logs into a special platform through data acquisition tools. Under the big data setting, many Internet enterprises utilize the offline analysis architecture based on Hadoop in order to reduce the cost of data format conversion and improve the efficiency of data acquisition. Examples include Facebook's open source tool Scribe, LinkedIn's open source tool Kafka, Taobao's open source tool Timetunnel, and Chukwa of Hadoop, etc. These tools can meet the demands of data acquisition and transmission with hundreds of MB per second.

5.3.2 Analysis at different levels

Big data analysis can also be classified into memory level analysis, Business Intelligence (BI) level analysis, and massive level analysis, which are examined in the following.

– *Memory-level analysis*: is for the case where the total data volume is smaller than the maximum memory of a cluster. Nowadays, the memory of server cluster surpasses hundreds of GB while even the TB level is common. Therefore, an internal database technology may be used, and hot data shall reside in the memory so as to improve the analytical efficiency. Memory-level analysis is extremely suitable for real-time analysis. MongoDB is a representative memory-level analytical architecture. With the development of SSD (Solid-State Drive), the capacity and performance of memory-level data analysis has been further improved and widely applied.

– *BI analysis*: is for the case when the data scale surpasses the memory level but may be imported into the BI analysis environment. The currently, mainstream BI products are provided with data analysis plans to support the level over TB.

– *Massive analysis*: is for the case when the data scale has completely surpassed the capacities of BI products and traditional relational databases. At present, most massive analysis utilize HDFS of Hadoop to store data and use MapReduce for data analysis. Most massive analysis belongs to the offline analysis category.

5.3.3 Analysis with different complexity

The time and space complexity of data analysis algorithms differ greatly from each other according to different kinds of data and application demands. For example, for applications that are amenable to parallel processing, a distributed algorithm may be designed and a parallel processing model may be used for data analysis.

5.4 Tools for big data mining and analysis

Many tools for big data mining and analysis are available, including professional and amateur software, expensive commercial software, and open source software. In this section, we briefly review the top five most widely used software, according to a survey of "What Analytics, Data mining, Big Data software that you used in the past 12 months for a real project?" of 798 professionals made by KDNuggets in 2012 [112].

- *R* (30.7 %): *R*, an open source programming language and software environment, is designed for data mining/analysis and visualization. While computing-intensive tasks are executed, code programmed with C, C++ and Fortran may be called in the R environment. In addition, skilled users can directly call *R* objects in C. Actually, *R* is a realization of the *S* language, which is an interpreted language developed by AT&T Bell Labs and used for data exploration, statistical analysis, and drawing plots. Compared to *S*, *R* is more popular since it is open source. *R* ranks top 1 in the KDNuggets 2012 survey. Furthermore, in a survey of "Design languages you have used for data mining/analysis in the past year" in 2012, *R* was also in the first place, defeating SQL and Java. Due to the popularity of *R*, database manufacturers, such as Teradata and Oracle, have released products supporting *R*.
- *Excel* (29.8 %): Excel, a core component of Microsoft Office, provides powerful data processing and statistical analysis capabilities. When Excel is installed, some advanced plug-ins, such as Analysis ToolPak and Solver Add-in, with powerful functions for data analysis are integrated initially, but such plug-ins can be used only if users enable them. Excel is also the only commercial software among the top five.
- *Rapid-I Rapidminer* (26.7 %): Rapidminer is an open source software used for data mining, machine learning, and predictive analysis. In an investigation of KDnuggets in 2011, it was more frequently used than R (ranked Top 1). Data mining and machine learning programs provided by RapidMiner include Extract, Transform and Load (ETL), data pre-processing and visualization, modeling, evaluation, and deployment. The data mining flow is described in XML and displayed through a graphic user interface (GUI). RapidMiner is written in Java. It integrates the learner and evaluation method of Weka, and works with R. Functions of Rapidminer are implemented with connection of processes including various operators. The entire flow can be deemed as a production line of a factory, with original data input and model results output. The operators can be considered as some specific functions with different input and output characteristics.
- *KNMINE* (21.8 %): KNIME (Konstanz Information Miner) is a user-friendly, intelligent, and open-source-rich data integration, data processing, data analysis, and data mining platform [113]. It allows users to create data flows or data channels in a visualized manner, to selectively run some or all analytical procedures, and provides analytical results, models, and interactive views. KNIME was written in Java and, based on Eclipse, provides more functions as plug-ins. Through plug-in files, users can insert processing modules for files, pictures, and time series, and integrate them into various open source projects, e.g., R and Weka. KNIME controls data integration, cleansing, conversion, filtering, statistics, mining, and finally data visualization. The entire development process is conducted under a visualized environment. KNIME is designed as a module-based and expandable framework. There is no dependence between its processing units and data containers, making it adaptive to the distributed environment and independent development. In addition, it is easy to expand KNIME. Developers can effortlessly expand various nodes and views of KNIME.
- *Weka/Pentaho* (14.8 %): Weka, abbreviated from Waikato Environment for Knowledge Analysis, is a free and open-source machine learning and data mining software written in Java. Weka provides such functions as data processing, feature selection, classification, regression, clustering, association rule, and visualization, etc. Pentaho is one of the most popular open-source BI software. It includes a web server platform and several tools to support reporting, analysis, charting, data integration, and data mining, etc., all aspects of BI. Weka's data processing algorithms are also integrated in Pentaho and can be directly called.

6 Big data applications

In the previous section, we examined big data analysis, which is the final and most important phase of the value chain of big data. Big data analysis can provide useful values via judgments, suggestions, supports, or decisions.

However, data analysis involves a wide range of applications, which frequently change and are extremely complex. In this section, we first review the evolution of data sources. We then examine six of the most important data analysis fields, including structured data analysis, text analysis, website analysis, multimedia analysis, network analysis, and mobile analysis. Finally, we introduce several key application fields of big data.

6.1 Application evolutions

Recently, big data analysis has been proposed as an advanced analytical technology, which typically includes large-scale and complex programs under specific analytical methods. As a matter of fact, data driven applications have emerged in the past decades. For example, as early as 1990s, BI has become a prevailing technology for business applications and, network search engines based on massive data mining processing emerged in the early 21st century. Some potential and influential applications from different fields and their data and analysis characteristics are discussed as follows.

- *Evolution of Commercial Applications*: The earliest business data was generally structured data, which was collected by companies from legacy systems and then stored in RDBMSs. Analytical techniques used in such systems were prevailing in the 1990s and was intuitive and simple, e.g., in the forms of reports, dashboard, queries with condition, search-based business intelligence, online transaction processing, interactive visualization, score cards, predictive modeling, and data mining [114]. Since the beginning of 21st century, networks and the World Wide Web (WWW) has been providing a unique opportunity for organizations to have online display and directly interact with customers. Abundant products and customer information, such as clickstream data logs and user behavior, can be acquired from the WWW. Product layout optimization, customer trade analysis, product suggestions, and market structure analysis can be conducted by text analysis and website mining techniques. As reported in [115], the quantity of mobile phones and tablet PC first surpassed that of laptops and PCs in 2011. Mobile phones and Internet of Things based on sensors are opening a new generation of innovation applications, and requiring considerably larger capacity of supporting location sensing, people oriented, and context-aware operation.
- *Evolution of Network Applications*: The early generation of the Internet mainly provided email and the WWW services. Text analysis, data mining, and webpage analysis have been applied to the mining of email contents and building search engines. Nowadays, most applications are web-based, regardless of their field and design goals. Network data accounts for a major percentage of the global data volume. Web has become a common platform for interconnected pages, full of various kinds of data, such as text, images, audio, videos, and interactive contents, etc. Therefore, a plentiful of advanced technologies used for semi-structured or unstructured data emerged at the right moment. For example, image analysis can extract useful information from images, (e.g., face recognition). Multimedia analysis technologies can be applied to automated video surveillance systems for business, law enforcement, and military applications. Since 2004, online social media, such as Internet forums, online communities, blogs, social networking services, and social multimedia websites, provide users with great opportunities to create, upload, and share contents.
- *Evolution of Scientific Applications*: Scientific research in many fields is acquiring massive data with high-throughput sensors and instruments, such as astrophysics, oceanology, genomics, and environmental research. The U.S. National Science Foundation (NSF) has recently announced the BIGDATA program to promote efforts to extract knowledge and insights from large and complex collections of digital data. Some scientific research disciplines have developed big data platforms and obtained useful outcomes. For example, in biology, iPlant [116] applies network infrastructure, physical computing resources, coordination environment, virtual machine resources, inter-operative analysis software, and data service to assist researchers, educators, and students in enriching plant sciences. The iPlant datasets have high varieties in form, including specification or reference data, experimental data, analog or model data, observation data, and other derived data.

6.2 Big data analysis fields

As discussed, we can divide data analysis research into six key technical fields, i.e., structured data analysis, text data analysis, web data analysis, multimedia data analysis, network data analysis, and mobile data analysis. Such a classification aims to emphasize data characteristics, but some of the fields may utilize similar basic technologies. Since data analysis has a broad scope and it is not easy to have a comprehensive coverage, we will focus on the key problems and technologies in data analysis in the following discussions.

6.2.1 Structured data analysis

Business applications and scientific research may generate massive structured data, of which the management and analysis rely on mature commercialized technologies, such as RDBMS, data warehouse, OLAP, and BPM (Business Process Management) [28]. Data analysis is mainly based on data mining and statistical analysis, both of which have been well studied over the past 30 years.

However, data analysis is still a very active research field and new application demands drive the development of new methods. For example, statistical machine learning based on exact mathematical models and powerful algorithms have been applied to anomaly detection [117] and energy control [118]. Exploiting data characteristics, time and space mining can extract knowledge structures hidden in high-speed data flows and sensors [119]. Driven by privacy protection in e-commerce, e-government, and health care applications, privacy protection data mining is an emerging research field [120]. Over the past decade, process mining is becoming a new research field especially in process analysis with event data [121].

6.2.2 Text data analysis

The most common format of information storage is text, e.g., emails, business documents, web pages, and social media. Therefore, text analysis is deemed to feature more business-based potential than structured data. Generally, text analysis is a process to extract useful information and knowledge from unstructured text. Text mining is inter-disciplinary, involving information retrieval, machine learning, statistics, computing linguistics, and data mining in particular. Most text mining systems are based on text expressions and natural language processing (NLP), with more emphasis on the latter. NLP allows computers to analyze, interpret, and even generate text. Some common NLP methods include lexical acquisition, word sense disambiguation, part-of-speech tagging, and probabilistic context free grammar [122]. Some NLP-based techniques have been applied to text mining, including information extraction, topic models, text summarization, classification, clustering, question answering, and opinion mining.

6.2.3 Web data analysis

Web data analysis has emerged as an active research field. It aims to automatically retrieve, extract, and evaluate information from Web documents and services so as to discover useful knowledge. Web analysis is related to several research fields, including database, information retrieval, NLP, and text mining. According to the different parts be mined, we classify Web data analysis into three related fields: Web content mining, Web structure mining, and Web usage mining [123].

Web content mining is the process to discover useful knowledge in Web pages, which generally involve several types of data, such as text, image, audio, video, code, metadata, and hyperlink. The research on image, audio, and video mining has recently been called multimedia analysis, which will be discussed in the Section 6.1.5. Since most Web content data is unstructured text data, the research on Web data analysis mainly centers around text and hypertext. Text mining is discussed in Section 6.1.3 while Hypertext mining involves the mining of the semi-structured HTML files that contain hyperlinks. Supervised learning and classification play important roles in hyperlink mining, e.g., email, newsgroup management, and Web catalogue maintenance [124]. Web content mining can be conducted with two methods: the information retrieval method and the database method. Information retrieval mainly assists in or improves information lookup, or filters user information according to deductions or configuration documents. The database method aims to simulate and integrate data in Web, so as to conduct more complex queries than searches based on key words.

Web structure mining involves models for discovering Web link structures. Here, the structure refers to the schematic diagrams linked in a website or among multiple websites. Models are built based on topological structures provided with hyperlinks with or without link description. Such models reveal the similarities and correlations among different websites and are used to classify website pages. Page Rank [125] and CLEVER [126] make full use of the models to look up relevant website pages. Topic-oriented crawler is another successful case by utilizing the models [127].

Web usage mining aims to mine auxiliary data generated by Web dialogues or activities. Web content mining and Web structure mining use the master Web data. Web usage data includes access logs at Web servers and proxy servers, browsers' history records, user profiles, registration data, user sessions or trades, cache, user queries, bookmark data, mouse clicks and scrolls, and any other kinds of data generated through interaction with the Web. As Web services and Web2.0 are becoming mature and popular, Web usage data will have increasingly high variety. Web usage mining plays key roles in personalized space, e-commerce, network privacy/security, and other emerging fields. For example, collaborative recommender systems can personalize e-commerce by utilizing the different preferences of users.

6.2.4 Multimedia data analysis

Multimedia data (mainly including images, audio, and videos) have been growing at an amazing speed, which is extracted useful knowledge and understand the semantemes by analysis. Because multimedia data is heterogeneous and most of such data contains richer information than simple structured data or text data, extracting information is confronted with the huge challenge of the semantic differences. Research on multimedia analysis covers many disciplines. Some recent research priorities include multimedia summarization, multimedia annotation, multimedia index and retrieval, multimedia suggestion, and multimedia event detection, etc.

Audio summarization can be accomplished by extracting the prominent words or phrases from metadata or synthesizing a new representation. Video summarization is to interpret the most important or representative video content sequence, and it can be static or dynamic. Static video summarization methods utilize a key frame sequence or context-sensitive key frames to represent a video. Such methods are simple and have been applied to many business applications (e.g., by Yahoo, AltaVista and Google), but their performance is poor. Dynamic summarization methods use a series of video frame to represent a video, and take other smooth measures to make the final summarization look more natural. In [128], the authors propose a topic-oriented multimedia summarization system (TOMS) that can automatically summarize the important information in a video belonging to a certain topic area, based on a given set of extracted features from the video.

Multimedia annotation inserts labels to describe contents of images and videos at both syntax and semantic levels. With such labels, the management, summarization, and retrieval of multimedia data can be easily implemented. Since manual annotation is both time and labor intensive, automatic annotation without any human interventions becomes highly appealing. The main challenge for automatic multimedia annotation is the semantic difference. Although much progress has been made, the performance of existing automatic annotation methods still needs to be improved. Currently, many efforts are being made to synchronously explore both manual and automatic multimedia annotation [129].

Multimedia indexing and retrieval involve describing, storing, and organizing multimedia information and assisting users to conveniently and quickly look up multimedia resources [130]. Generally, multimedia indexing and retrieval include five procedures: structural analysis, feature extraction, data mining, classification and annotation, query and retrieval [131]. Structural analysis aims to segment a video into several semantic structural elements, including lens boundary detection, key frame extraction, and scene segmentation, etc. According to the result of structural analysis, the second procedure is feature extraction, which mainly includes further mining the features of key frames, objects, texts, and movements, which are the foundation of video indexing and retrieval. Data mining, classification, and annotation are to utilize the extracted features to find the modes of video contents and put videos into scheduled categories so as to generate video indexes. Upon receiving a query, the system will use a similarity measurement method to look up a candidate video. The retrieval result optimizes the related feedback.

Multimedia recommendation is to recommend specific multimedia contents according to users' preferences. It is proven to be an effective approach to provide personalized services. Most existing recommendation systems can be classified into content-based systems and collaborative-filtering-based systems. The content-based methods identify general features of users or their interesting, and recommend users for other contents with similar features. These methods largely rely on content similarity measurement, but most of them are troubled by analysis limitation and excessive specifications. The collaborative-filtering-based methods identify groups with similar interests and recommend contents for group members according to their behavior [132]. Presently, a mixed method is introduced, which integrates advantages of the aforementioned two types of methods to improve recommendation quality [133].

The U.S. National Institute of Standards and Technology (NIST) initiated the TREC Video Retrieval Evaluation for detecting the occurrence of an event in video-clips based on Event Kit, which contains some text description related to concepts and video examples [134]. In [135], the author proposed a new algorithm on special multimedia event detection using a few positive training examples. The research on video event detection is still in its infancy, and mainly focuses on sports or news events, running or abnormal events in monitoring videos, and other similar events with repetitive patterns.

6.2.5 Network data analysis

Network data analysis evolved from the initial quantitative analysis [136] and sociological network analysis [137] into the emerging online social network analysis in the beginning of 21st century. Many online social networking services include Twitter, Facebook, and LinkedIn, etc. have become increasingly popular over the years. Such online social network services generally include massive linked data and content data. The linked data is mainly in the form of graphic structures, describing the communications between two entities. The content data contains text, image, and other network multimedia data. The rich content in such networks brings about both unprecedented challenges and opportunities for data analysis. In accordance with the data-centered perspective, the existing research on social networking service contexts can be classified into two categories: link-based structural analysis and content-based analysis [138].

The research on link-based structural analysis has always been committed on link prediction, community discovery, social network evolution, and social influence analysis, etc. SNS may be visualized as graphs, in which every vertex corresponds to a user and edges correspond to the correlations among users. Since SNS are dynamic networks, new vertexes and edges are continually added to the graphs. Link prediction is to predict the possibility of future connection between two vertexes. Many techniques can be used for link prediction, e.g., feature-based classification, probabilistic methods, and Linear Algebra. Feature-based classification is to select a group of features for a vertex and utilize the existing link information to generate binary classifiers to predict the future link [139]. Probabilistic methods aim to build models for connection probabilities among vertexes in SNS [140]. Linear Algebra computes the similarity between two vertexes according to the singular similar matrix [141]. A community is represented by a subgraphic matrix, in which edges connecting vertexes in the sub-graph feature high density while the edges between two sub-graphs feature much lower density [142].

Many methods for community detection have been proposed and studied, most of which are topology-based target functions relying on the concept of capturing community structure. Du et al. utilized the property of overlapping communities in real life to propose an effective large-scale SNS community detection method [143]. The research on SNS aims to look for a law and deduction model to interpret network evolution. Some empirical studies found that proximity bias, geographical limitations, and other factors play important roles in SNS evolution [144–146], and some generation methods are proposed to assist network and system design [147].

Social influence refers to the case when individuals change their behavior under the influence of others. The strength of social influence depends on the relation among individuals, network distances, time effect, and characteristics of networks and individuals, etc. Marketing, advertisement, recommendation, and other applications can benefit from social influence by qualitatively and quantitatively measuring the influence of individuals on others [148, 149]. Generally, if the proliferation of contents in SNS is considered, the performance of link-based structural analysis may be further improved.

Content-based analysis in SNS is also known as social media analysis. Social media include text, multimedia, positioning, and comments. However, social media analysis is confronted with unprecedented challenges. First, massive and continually growing social media data should be automatically analyzed within a reasonable time window. Second, social media data contains much noise. For example, blogosphere contains a large number of spam blogs, and so does trivial Tweets in Twitter. Third, SNS are dynamic networks, which are frequently and quickly varying and updated. The existing research on social media analysis is still in its infancy. Considering that SNS contains massive information, transfer learning in heterogeneous networks aims to transfer knowledge information among different media [150].

6.2.6 Mobile data analysis

By April 2013, Android Apps has provided more than 650,000 applications, covering nearly all categories. By the end of 2012, the monthly mobile data flow has reached 885 PB [151]. The massive data and abundant applications call for mobile analysis, but also bring about a few challenges. As a whole, mobile data has unique characteristics, e.g., mobile sensing, moving flexibility, noise, and a large amount of redundancy. Recently, new research on mobile analysis has been started in different fields. Since the research on mobile analysis is just started, we will only introduce some recent and representative analysis applications in this section.

With the growth of numbers of mobile users and improved performance, mobile phones are now useful for building and maintaining communities, such as communities with geographical locations and communities based on different cultural backgrounds and interests(e.g., the latest Webchat). Traditional network communities or SNS communities are in short of online interaction among members, and the communities are active only when members are sitting before computers. On the contrary, mobile phones can support rich interaction at any time and anywhere. Mobile communities are defined as that a group of individuals with the same hobbies (i.e., health, safety, and entertainment, etc.) gather together on networks, meet to make a common goal, decide measures through consultation to achieve the goal, and start to implement their plan [152]. In [153], the authors proposed a qualitative model of a mobile community. It is now widely believed that mobile community applications will greatly promote the development of the mobile industry.

Recently, the progress in wireless sensor, mobile communication technology, and stream processing enable people to build a body area network to have real-time monitoring of people's health. Generally, medical data from various sensors have different characteristics in terms of attributes, time and space relations, as well as physiological relations, etc.

In addition, such datasets involve privacy and safety protection. In [154], Garg et al. introduce a multi-modal transport analysis mechanism of raw data for real-time monitoring of health. Under the circumstance that only highly comprehensive characteristics related to health are available, Park et al. in [155] examined approaches to better utilize.

Researchers from Gjovik University College in Norway and Derawi Biometrics collaborated to develop an application for smart phones, which analyzes paces when people walk and uses the pace information for unlocking the safety system [11]. In the meanwhile, Robert Delano and Brian Parise from Georgia Institute of Technology developed an application called iTrem, which monitors human body trembling with a built-in seismograph in a mobile phone, so as to cope with Parkinson and other nervous system diseases [11].

6.3 Key applications of big data

6.3.1 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. 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 building a customer drop out warning model, the bank can sell high-yield financial products to the top 20 % customers who are most likely to drop out so as to retain them. As a result, the drop out ratios of customers with Gold Cards and Sunflower Cards have been reduced by 15 % and 7 %, respectively. 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.

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. Data Cube of Taobao is a big data application on the Taobao platform, through which, merchants can be ware 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.

6.3.2 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. In 2011, UPS drivers have driven for nearly 48.28 million km less.

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. The application of smart city brings about benefits in many aspects for Dade County. For instance, the Department of Park Management of Dade County saved one million USD in water bills due to timely identifying and fixing water pipes that were running and leaking this year.

6.3.3 Application of online social network-oriented big data

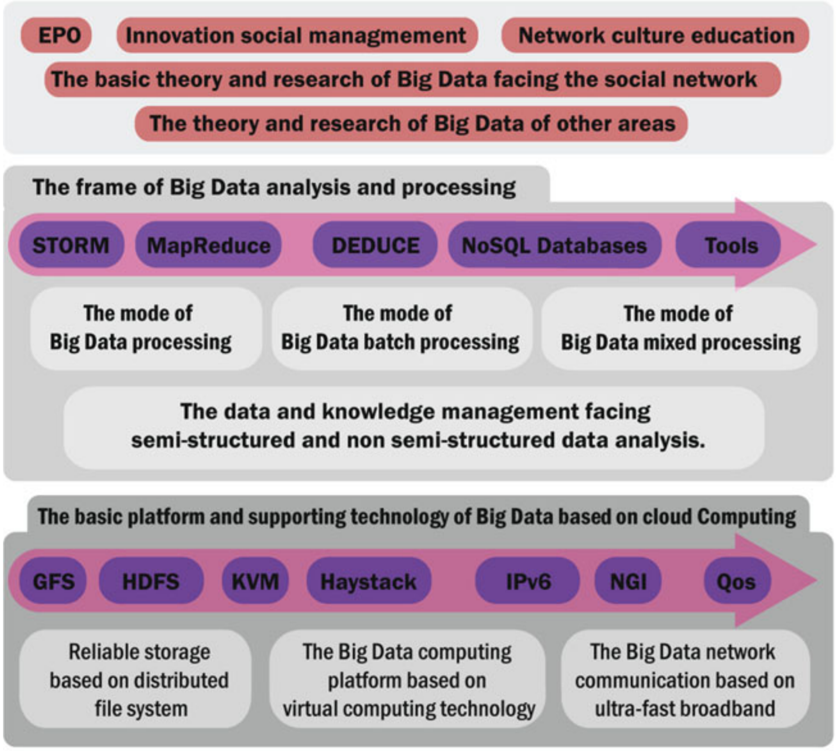
Online SNS is a social structure constituted by social individuals and connections among individuals based on an information network. Big data of online SNS mainly comes from instant messages, online social, micro blog, and shared space, etc, which represents various user activities. The analysis of 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. Fig. 5 illustrates the technical framework of the application of big data of online SNS. 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.

- *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.
- *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.

The U.S. Santa Cruz Police Department experimented by applying data for predictive analysis. By analyzing SNS, the police department can discover crime trends and crime modes, and even predict the crime rates in major regions [11].

In April 2013, Wolfram Alpha, a computing and search engine company, studied the law of social behavior by analyzing social data of more than one million American users of Facebook. According to the analysis, it was found that most Facebook users fall in love in their early 20s, and get engaged when they are about 27 years old, then get married when they are about 30 years old. Finally, their marriage relationships exhibit slow changes between 30 and 60 years old. Such research results are highly consistent with the demographic census data of the U.S. In addition, Global Pulse conducted a research that revealed some laws in social and economic activities using SNS data. This project utilized publicly available Twitter messages in English, Japanese, and Indonesian from July 2010 to October 2011, to analyze topics related to food, fuel, housing, and loan. The goal is to better understand public behavior and concerns. This project analyzed SNS big data from several aspects: 1) predicting the occurrence of abnormal events by detecting the sharp growth

Fig. 5 Enabling technologies for online social network-oriented big data


EPO
Innovation social managment
Network culture education
The basic theory and research of Big Data facing the social network
The theory and research of Big Data of other areas
The frame of Big Data analysis and processing
STORM
MapReduce
DEDUCE
NoSQL Databases
Tools
The mode of Big Data processing
The mode of Big Data batch processing
The mode of Big Data mixed processing
The data and knowledge management facing semi-structured and non semi-structured data analysis.
The basic platform and supporting technology of Big Data based on cloud Computing
GFS
HDFS
KVM
Haystack
IPv6
NGI
Qos
Reliable storage based on distributed file system
The Big Data computing platform based on virtual computing technology
The Big Data network communication based on ultra-fast broadband

or drop of the amount of topics, 2) observing the weekly and monthly trends of dialogs on Twitter; developing models for the variation in the level of attention on specific topics over time, 3) understanding the transformation trends of user behavior or interest by comparing ratios of different sub-topics, and 4) predicting trends with external indicators involved in Twitter dialogues. As a classic example, the project discovered that the change of food price inflation from the official statistics of Indonesia matches the number of Tweets to rice price on Twitter, as shown in Fig. 6.

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:

- *Early Warning*: to rapidly cope with crisis if any by detecting abnormalities in the usage of electronic devices and services.
- *Real-time Monitoring*: to provide accurate information for the formulation of policies and plans by monitoring the current behavior, emotion, and preference of users.
- *Real-time Feedback*: acquire groups' feedbacks against some social activities based on real-time monitoring.

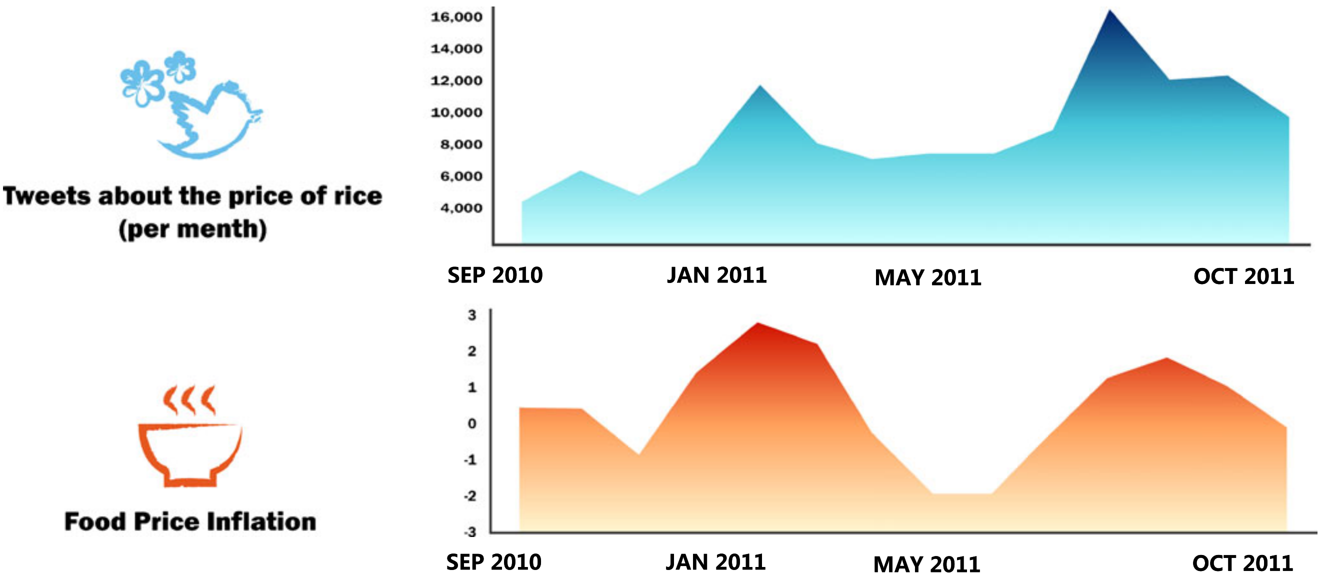
6.3.4 Applications of healthcare and medical big data

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.

The Mount Sinai Medical Center in the U.S. utilizes technologies of Ayasdi, a big data company, to analyze all genetic sequences of Escherichia Coli, including over one million DNA variants, to investigate why bacterial strains resist antibiotics. Ayasdi's uses topological data analysis, a brand-new mathematic research method, to understand data characteristics.

HealthVault of Microsoft, launched in 2007, is an excellent application of medical big data launched in 2007. Its goal is to manage individual health information in individual and family medical devices. Presently, health information can be entered and uploaded with mobile smart devices and



[URL] http://www.unglobalpulse.org/projects/twitter-and-perceptions-crisis-related-stress

Fig. 6 The correlation between Tweets about rice price and food price inflation

imported from individual medical records by a third-party agency. In addition, it can be integrated with a third-party application with the software development kit (SDK) and open interface.

6.3.5 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.

6.3.6 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, such as (i) power utilization habits of users, (ii) phasor measurement data, which are measured by phasor measurement unit (PMU) deployed national-wide, (iii) energy consumption data measured by the smart meters in the Advanced Metering Infrastructure (AMI), (iv) energy market pricing and bidding data, (v) 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.

- *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. For example, researchers from University of California, Los Angeles designed an “electric map” according to the big data theory and made a California map by integrating census information and real-time power utilization information provided by electric power companies. The map takes a block as a unit to demonstrate the power consumption of every block at the moment. It can even compare the power consumption of the block with the average income per capita and building types, so as to reveal more accurate power usage habits of all kinds of groups in the community. This map provides effective and visual load forecast for power grid planning in a city. 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.
- *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.
- *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.

7 Conclusion, open issues, and outlook

In this paper, we review the background and state-of-the-art of big data. Firstly, we introduce the general background of big data and review related technologies, such as could computing, IoT, data centers, and Hadoop. Then we focus on the four phases of the value chain of big data, i.e., data generation, data acquisition, data storage, and data analysis. For each phase, we introduce the general background, discuss the technical challenges, and review the latest advances. We finally reviewed 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.

In the remainder of this section, 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 early stage. Considerable research efforts are needed to improve the efficiency of display, storage, and analysis of big data.

7.1.1 Theoretical research

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

- *Fundamental problems of big data*: There is a compelling need for a rigorous and holistic definition of big data, a structural model of big data, a formal description of big data, and a theoretical system of data science. 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 the existing models are not strictly verified.
- *Standardization of big data*: An evaluation system of data quality and an evaluation standard/benchmark 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 methods. The performance can only be evaluated when the system is implemented and deployed, which could not horizontally compare advantages and disadvantages of various alternative solutions even before and after the implementation 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 and rigorously evaluate data quality.
- *Evolution of big data computing modes*: This includes memory mode, data flow mode, PRAM mode, and MR mode, etc. The emergence of big data triggers the advances of algorithm design, which has been 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 of big data*: Due to wide and diverse 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 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 of big data*: The real-time performance of big data is also a key problem in many application scenarios. Effective means to define the life cycle of data, compute the rate of depreciation of data, and build computing models of real-time and online applications, will influence the analysis results of big data.
- *Processing of big data*: As big data research is advanced, new problems on big data processing arise from the traditional data analysis, including (i) data re-utilization, with the increase of data scale, more values may be mined from re-utilization of existing data; (ii) data re-organization, datasets in different businesses can be re-organized, which can be mined more value; (iii) data exhaust, which means wrong data during acquisition. In big data, not only the correct data but also the wrong data should be utilized to generate more value.

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 big data oriented database and Internet technologies, storage models and databases suitable for 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 big data. Related problems include 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 comprehensive utilization of multiple datasets is far higher than the sum value of individual dataset. Therefore, the integration of different data sources is a timely problem. Data integration is confronted with many challenges, such as different data patterns and a large amount of redundant data. Data provenance is the process of data generation and evolution over time, and mainly used to investigate multiple datasets other than a single dataset. Therefore, it is worth studying 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 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: Big data privacy includes two aspects: (i) Protection of personal privacy during data acquisition: personal interests, habits, and body properties, etc. of users may be more easily acquired, and users may not be aware. (ii) Personal privacy data may also be leaked during storage, transmission, and usage, even if acquired with the permission of users. For example, Facebook is deemed as a big data company with the most SNS data currently. According to a report [156], Ron Bowes, a researcher of Skull Security, acquired data in the public pages of Facebook users who fail to modify their privacy setting via an information acquisition tool. Ron Bowes packaged such data into a 2.8 GB package and created a BitTorrent (BT) seed for others to download. The analysis capacity of big data may lead to privacy mining from seemingly simple information. Therefore, privacy protection will become a new and challenging problem.
- Data quality: Data quality influences big data utilization. Low quality data wastes transmission and storage resources with poor usability. There are a lot of factors that may restrict data quality, for example, generation, acquisition, 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 related problems have not been well addressed yet. Therefore, effective methods to automatically detect data quality and repair some damaged data need to be investigated.

- Big data safety mechanism: Big data brings about challenges to data encryption due to its large scale and high diversity. The performance of previous encryption methods on small and medium-scale data could not meet the demands of big data, so 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 about 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 analysis of 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.

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 credibility, backup and recovery, completeness maintenance, and security 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 drives the development of data. In the big data era, with the prominence of data value and advances in "I" (Information), data will drive the progress of technologies in the near future. Big data will not only have the social and economic impact, but also influence everyone's ways of living and thinking, which is just happening. We could not predict the future but may take precautions for possible events to occur in the future.

- *Data with a larger scale, higher diversity, and more complex structures*: Although technologies represented by Hadoop have achieved a great success, such technologies are expected to fall behind and will be replaced given the rapid development of big data. The theoretical basis of Hadoop has emerged as early as 2006. Many researchers have concerned better ways to cope with larger-scale, higher diversity, and more complexly structured data. These efforts are represented by (Globally-Distributed Database) Spanner of Google and fault-tolerant, expandable, distributed relational database F1. 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.
- *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. 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, IoT, 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 terms of data acquisition, storage, processing, analysis, and information security, etc. Then, impacts of big data on production management, business operation and decision making, etc., shall be examined for modern enterprises from the management perspective. Moreover, 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., as in text and image editors. In big data applications, mixed data is very useful for decision making. Only when the analytical results are friendly displayed, it may be effectively 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 consist of data structures and algorithms, and data structures are used to store data. 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 process-oriented. As business data is becoming more complex, object-oriented design methods are developed. Nowadays, the complexity of business data has far surpassed business logic. Consequently, programs are gradually transformed from algorithm-intensive to data-intensive. It is anticipated that 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 triggers the revolution of thinking*: Gradually, big data and its analysis will profoundly influence our ways of thinking. In [11], the authors summarize the thinking revolution triggered by big data as follows:
  - During data analysis, we will try to utilize all data other than only analyzing a small set of 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."

Throughout the history of human society, the demands and willingness of human beings are always the source powers to promote scientific and technological progress. Big data may provides reference answers for human beings to make decisions through mining and analytical processing, but it 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 the human brain. 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.

Acknowledgments This work was supported by China National Natural Science Foundation (No. 61300224), the Ministry of Science and Technology (MOST), China, the International Science and Technology Collaboration Program (Project No.:2014DFT10070), and the Hubei Provincial Key Project (No. 2013CFA051). Shiwen Mao's research is supported in part by the US NSF under grants CNS-1320664, CNS-1247955, and CNS-0953513, and through the NSF Broadband Wireless Access & Applications Center (BWAC) site at Auburn University.

References

1. Gantz J, Reinsel D (2011) Extracting value from chaos. IDC iView, pp 1–12
2. Fact sheet: Big data across the federal government (2012). http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_3_29_2012.pdf
3. Cukier K (2010) Data, data everywhere: a special report on managing information. Economist Newspaper
4. Drowning in numbers - digital data will flood the planet- and help us understand it better (2011). http://www.economist.com/blogs/dailychart/2011/11/bigdata-0
5. Lohr S (2012) The age of big data. New York Times, pp 11
6. Yuki N (2011) Following digital breadcrumbs to big data gold. http://www.npr.org/2011/11/29/142521910/thedigitalbreadcrumbs-that-lead-to-big-data
7. Yuki N The search for analysts to make sense of big data (2011). http://www.npr.org/2011/11/30/142893065/the-searchforanalysts-to-make-sense-of-big-data
8. Big data (2008). http://www.nature.com/news/specials/bigdata/index.html
9. Special online collection: dealing with big data (2011). http://www.sciencemag.org/site/special/data/
10. Manyika J, McKinsey Global Institute, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2011) Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute
11. Mayer-Schönberger V, Cukier K (2013) Big data: a revolution that will transform how we live, work, and think. Eamon Dolan/Houghton Mifflin Harcourt
12. Laney D (2001) 3-d data management: controlling data volume, velocity and variety. META Group Research Note, 6 February
13. Zikopoulos P, Eaton C et al (2011) Understanding big data: analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media
14. Meijer E (2011) The world according to linq. Communications of the ACM 54(10):45–51
15. Beyer M (2011) Gartner says solving big data challenge involves more than just managing volumes of data. Gartner. http://www.gartner.com/it/page.jsp
16. O. R. Team (2011) Big data now: current perspectives from OReilly Radar. OReilly Media
17. Grobelnik M (2012) Big data tutorial. http://videolectures.net/eswc2012grobelnikbigdata/
18. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2008) Detecting influenza epidemics using search engine query data. Nature 457(7232):1012–1014
19. DeWitt D, Gray J (1992) Parallel database systems: the future of high performance database systems. Commun ACM 35(6):85–98

20. Walter T (2009) Teradata past, present, and future. UCI ISG lecture series on scalable data management
21. Ghemawat S, Gobioff H, Leung S-T (2003) The google file system. In: ACM SIGOPS Operating Systems Review, vol 37. ACM, pp 29–43
22. Dean J, Ghemawat S (2008) Mapreduce: simplified data processing on large clusters. Commun ACM 51(1):107–113
23. Hey AJG, Tansley S, Tolle KM et al (2009) The fourth paradigm: data-intensive scientific discovery
24. Howard JH, Kazar ML, Menees SG, Nichols DA, Satyanarayanan M, Sidebotham RN, West MJ (1988) Scale and performance in a distributed file system. ACM Trans Comput Syst (TOCS) 6(1):51–81
25. Cattell R (2011) Scalable sql and nosql data stores. ACM SIGMOD Record 39(4):12–27
26. Labrinidis A, Jagadish HV (2012) Challenges and opportunities with big data. Proc VLDB Endowment 5(12):2032–2033
27. Chaudhuri S, Dayal U, Narasayya V (2011) An overview of business intelligence technology. Commun ACM 54(8): 88–98
28. Agrawal D, Bernstein P, Bertino E, Davidson S, Dayal U, Franklin M, Gehrke J, Haas L, Halevy A, Han J et al (2012) Challenges and opportunities with big data. A community white paper developed by leading researches across the United States
29. Sun Y, Chen M, Liu B, Mao S (2013) Far: a fault-avoidant routing method for data center networks with regular topology. In: Proceedings of ACM/IEEE symposium on architectures for networking and communications systems (ANCS'13). ACM
30. Wiki (2013). Applications and organizations using hadoop. http://wiki.apache.org/hadoop/PoweredBy
31. Bahga A, Madisetti VK (2012) Analyzing massive machine maintenance data in a computing cloud. IEEE Transac Parallel Distrib Syst 23(10):1831–1843
32. Gunarathne T, Wu T-L, Choi JY, Bae S-H, Qiu J (2011) Cloud computing paradigms for pleasingly parallel biomedical applications. Concurr Comput Prac Experience 23(17):2338–2354
33. Gantz J, Reinsel D (2010) The digital universe decade-are you ready. External publication of IDC (Analyse the Future) information and data, pp 1–16
34. Bryant RE (2011) Data-intensive scalable computing for scientific applications. Comput Sci Eng 13(6):25–33
35. Wahab MHA, Mohd MNH, Hanafi HF, Mohsin MFM (2008) Data pre-processing on web server logs for generalized association rules mining algorithm. World Acad Sci Eng Technol 48:2008
36. Nanopoulos A, Manolopoulos Y, Zakrzewicz M, Morzy T (2002) Indexing web access-logs for pattern queries. In: Proceedings of the 4th international workshop on web information and data management. ACM, pp 63–68
37. Joshi KP, Joshi A, Yesha Y (2003) On using a warehouse to analyze web logs. Distrib Parallel Databases 13(2):161–180
38. Chandramohan V, Christensen K (2002) A first look at wired sensor networks for video surveillance systems. In: Proceedings LCN 2002, 27th annual IEEE conference on local computer networks. IEEE, pp 728–729
39. Selavo L, Wood A, Cao Q, Sookoor T, Liu H, Srinivasan A, Wu Y, Kang W, Stankovic J, Young D et al (2007) Luster: wireless sensor network for environmental research. In: Proceedings of the 5th international conference on Embedded networked sensor systems. ACM, pp 103–116
40. Barrenetxea G, Ingelrest F, Schaefer G, Vetterli M, Couach O, Parlange M (2008) Sensorscope: out-of-the-box environmental monitoring. In: Information processing in sensor networks, 2008, international conference on IPSN'08. IEEE, pp 332–343
41. Kim Y, Schmid T, Charbiwala ZM, Friedman J, Srivastava MB (2008) Nawms: nonintrusive autonomous water monitoring system. In: Proceedings of the 6th ACM conference on Embedded network sensor systems. ACM, pp 309–322
42. Kim S, Pakzad S, Culler D, Demmel J, Fenves G, Glaser S, Turon M (2007) Health monitoring of civil infrastructures using wireless sensor networks. In Information Processing in Sensor Networks 2007, 6th International Symposium on IPSN 2007. IEEE, pp 254–263
43. Ceriotti M, Mottola L, Picco GP, Murphy AL, Guna S, Corra M, Pozzi M, Zonta D, Zanon P (2009) Monitoring heritage buildings with wireless sensor networks: the torre aquila deployment. In: Proceedings of the 2009 International Conference on Information Processing in Sensor Networks. IEEE Computer Society, pp 277–288
44. Tolle G, Polastre J, Szewczyk R, Culler D, Turner N, Tu K, Burgess S, Dawson T, Buonadonna P, Gay D et al (2005) A macroscope in the redwoods. In: Proceedings of the 3rd international conference on embedded networked sensor systems. ACM, pp 51–63
45. Wang F, Liu J (2011) Networked wireless sensor data collection: issues, challenges, and approaches. IEEE Commun Surv Tutor 13(4):673–687
46. Cho J, Garcia-Molina H (2002) Parallel crawlers. In: Proceedings of the 11th international conference on World Wide Web. ACM, pp 124–135
47. Choudhary S, Dincturk ME, Mirtaheri SM, Moosavi A, von Bochmann G, Jourdan G-V, Onut I-V (2012) Crawling rich internet applications: the state of the art. In: CASCON. pp 146–160
48. Ghani N, Dixit S, Wang T-S (2000) On ip-over-wdm integration. IEEE Commun Mag 38(3):72–84
49. Manchester J, Anderson J, Doshi B, Dravida S, Ip over sonet (1998) IEEE Commun Mag 36(5):136–142
50. Jinno M, Takara H, Kozicki B (2009) Dynamic optical mesh networks: drivers, challenges and solutions for the future. In: Optical communication, 2009, 35th European conference on ECOC'09. IEEE, pp 1–4
51. Barroso LA, Hölzle U (2009) The datacenter as a computer: an introduction to the design of warehouse-scale machines. Synt Lect Comput Archit 4(1):1–108
52. Armstrong J (2009) Ofdm for optical communications. J Light Technol 27(3):189–204
53. Shieh W (2011) Ofdm for flexible high-speed optical networks. J Light Technol 29(10):1560–1577
54. Cisco data center interconnect design and deployment guide (2010)
55. Greenberg A, Hamilton JR, Jain N, Kandula S, Kim C, Lahiri P, Maltz DA, Patel P, Sengupta S (2009) Vl2: a scalable and flexible data center network. In ACM SIGCOMM computer communication review, vol 39. ACM, pp 51–62
56. Guo C, Lu G, Li D, Wu H, Zhang X, Shi Y, Tian C, Zhang Y, Lu S (2009) Bcube: a high performance, server-centric network architecture for modular data centers. ACM SIGCOMM Comput Commun Rev 39(4):63–74
57. Farrington N, Porter G, Radhakrishnan S, Bazzaz HH, Subramanya V, Fainman Y, Papen G, Vahdat A (2011) Helios: a hybrid electrical/optical switch architecture for modular data centers. ACM SIGCOMM Comput Commun Rev 41(4):339–350
58. Abu-Libdeh H, Costa P, Rowstron A, O'Shea G, Donnelly A (2010) Symbiotic routing in future data centers. ACM SIGCOMM Comput Commun Rev 40(4):51–62
59. Lam C, Liu H, Koley B, Zhao X, Kamalov V, Gill V, Fiber optic communication technologies: what's needed for datacenter network operations (2010) IEEE Commun Mag 48(7):32–39

60. Wang G, Andersen DG, Kaminsky M, Papagiannaki K, Ng TS, Kozuch M, Ryan M (2010) c-through: Part-time optics in data centers. In: ACM SIGCOMM Computer Communication Review, vol 40. ACM, pp 327–338
61. Ye X, Yin Y, Yoo SJB, Mejia P, Proietti R, Akella V (2010) Dos: a scalable optical switch for datacenters. In Proceedings of the 6th ACM/IEEE symposium on architectures for networking and communications systems. ACM, p 24
62. Singla A, Singh A, Ramachandran K, Xu L, Zhang Y (2010) Proteus: a topology malleable data center network. In Proceedings of the 9th ACM SIGCOMM workshop on hot topics in networks. ACM, p 8
63. Liboiron-Ladouceur O, Cerutti I, Raponi PG, Andriolli N, Castoldi P (2011) Energy-efficient design of a scalable optical multiplane interconnection architecture. IEEE J Sel Top Quantum Electron 17(2):377–383
64. Kodi AK, Louri A (2011) Energy-efficient and bandwidth-reconfigurable photonic networks for high-performance computing (hpc) systems. IEEE J Sel Top Quantum Electron 17(2):384–395
65. Zhou X, Zhang Z, Zhu Y, Li Y, Kumar S, Vahdat A, Zhao BY, Zheng H (2012) Mirror mirror on the ceiling: flexible wireless links for data centers. ACM SIGCOMM Comput Commun Rev 42(4):443–454
66. Lenzerini M (2002) Data integration: a theoretical perspective. In: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems. ACM, pp 233–246
67. Cafarella MJ, Halevy A, Khoussainova N (2009) Data integration for the relational web. Proc VLDB Endowment 2(1):1090–1101
68. Maletic JI, Marcus A (2000) Data cleansing: beyond integrity analysis. In: IQ. Citeseer, pp 200–209
69. Kohavi R, Mason L, Parekh R, Zheng Z (2004) Lessons and challenges from mining retail e-commerce data. Mach Learn 57(1-2):83–113
70. Chen H, Ku W-S, Wang H, Sun M-T (2010) Leveraging spatio-temporal redundancy for rfid data cleansing. In: Proceedings of the 2010 ACM SIGMOD international conference on management of data. ACM, pp 51–62
71. Zhao Z, Ng W (2012) A model-based approach for rfid data stream cleansing. In Proceedings of the 21st ACM international conference on information and knowledge management. ACM, pp 862–871
72. Khoussainova N, Balazinska M, Suciu D (2008) Probabilistic event extraction from rfid data. In: Data Engineering, 2008. IEEE 24th international conference on ICDE 2008. IEEE, pp 1480–1482
73. Herbert KG, Wang JTL (2007) Biological data cleaning: a case study. Int J Inf Qual 1(1):60–82
74. Tsai T-H, Lin C-Y (2012) Exploring contextual redundancy in improving object-based video coding for video sensor networks surveillance. IEEE Transac Multmed 14(3):669–682
75. Sarawagi S, Bhamidipaty A (2002) Interactive deduplication using active learning. In Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 269–278
76. Kamath U, Compton J, Dogan RI, Jong KD, Shehu A (2012) An evolutionary algorithm approach for feature generation from sequence data and its application to dna splice site prediction. IEEE/ACM Transac Comput Biol Bioinforma (TCBB) 9(5):1387–1398
77. Leung K-S, Lee KH, Wang J-F, Ng EYT, Chan HLY, Tsui SKW, Mok TSK, Tse PC-H, Sung JJ-Y (2011) Data mining on dna sequences of hepatitis b virus. IEEE/ACM Transac Comput Biol Bioinforma 8(2):428–440
78. Huang Z, Shen H, Liu J, Zhou X (2011) Effective data co-reduction for multimedia similarity search. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. ACM, pp 1021–1032
79. Bleiholder J, Naumann F (2008) Data fusion. ACM Comput Surv (CSUR) 41(1):1
80. Brewer EA (2000) Towards robust distributed systems. In: PODC. p 7
81. Gilbert S, Lynch N (2002) Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services. ACM SIGACT News 33(2):51–59
82. McKusick MK, Quinlan S (2009) Gfs: eqvolution on fast-forward. ACM Queue 7(7):10
83. Chaiken R, Jenkins B, Larson P-Å, Ramsey B, Shakib D, Weaver S, Zhou J (2008) Scope: easy and efficient parallel processing of massive data sets. Proc VLDB Endowment 1(2):1265–1276
84. Beaver D, Kumar S, Li HC, Sobel J, Vajgel P et al (2010) Finding a needle in haystack: facebook's photo storage. In OSDI, vol 10. pp 1–8
85. DeCandia G, Hastorun D, Jampani M, Kakulapati G, Lakshman A, Pilchin A, Sivasubramanian S, Vosshall P, Vogels W (2007) Dynamo: amazon's highly available key-value store. In: SOSP, vol 7. pp 205–220
86. Karger D, Lehman E, Leighton T, Panigrahy R, Levine M, Lewin D (1997) Consistent hashing and random trees: distributed caching protocols for relieving hot spots on the world wide web. In: Proceedings of the twenty-ninth annual ACM symposium on theory of computing. ACM, pp 654–663
87. Chang F, Dean J, Ghemawat S, Hsieh WC, Wallach DA, Burrows M, Chandra T, Fikes A, Gruber RE (2008) Bigtable: a distributed storage system for structured data. ACM Trans Comput Syst (TOCS) 26(2):4
88. Burrows M (2006) The chubby lock service for loosely-coupled distributed systems. In: Proceedings of the 7th symposium on Operating systems design and implementation. USENIX Association, pp 335–350
89. Lakshman A, Malik P (2009) Cassandra: structured storage system on a p2p network. In: Proceedings of the 28th ACM symposium on principles of distributed computing. ACM, pp 5–5
90. George L (2011) HBase: the definitive guide. O'Reilly Media Inc
91. Judd D (2008) hypertable-0.9. 0.4-alpha
92. Chodorow K (2013) MongoDB: the definitive guide. O'Reilly Media Inc
93. Crockford D (2006) The application/json media type for javascript object notation (json)
94. Murty J (2009) Programming amazon web services: S3, EC2, SQS, FPS, and SimpleDB. O'Reilly Media Inc
95. Anderson JC, Lehnardt J, Slater N (2010) CouchDB: the definitive guide. O'Reilly Media Inc
96. Blanas S, Patel JM, Ercegovac V, Rao J, Shekita EJ, Tian Y (2010) A comparison of join algorithms for log processing in mapreduce. In: Proceedings of the 2010 ACM SIGMOD international conference on management of data. ACM, pp 975–986
97. Yang H-C, Parker DS (2009) Traverse: simplified indexing on large map-reduce-merge clusters. In: Database systems for advanced applications. Springer, pp 308–322
98. Pike R, Dorward S, Griesemer R, Quinlan S (2005) Interpreting the data: parallel analysis with sawzall. Sci Program 13(4):277–298
99. Gates AF, Natkovich O, Chopra S, Kamath P, Narayanamurthy SM, Olston C, Reed B, Srinivasan S, Srivastava U (2009) Building a high-level dataflow system on top of map-reduce: the pig experience. Proceedings VLDB Endowment 2(2):1414–1425

100. Thusoo A, Sarma JS, Jain N, Shao Z, Chakka P, Anthony S, Liu H, Wyckoff P, Murthy R (2009) Hive: a warehousing solution over a map-reduce framework. Proc VLDB Endowment 2(2):1626–1629
101. Isard M, Budiu M, Yu Y, Birrell A, Fetterly D (2007) Dryad: distributed data-parallel programs from sequential building blocks. ACM SIGOPS Oper Syst Rev 41(3):59–72
102. Yu Y, Isard M, Fetterly D, Budiu M, Erlingsson Ú, Gunda PK, Currey J (2008) Dryadlinq: a system for general-purpose distributed data-parallel computing using a high-level language. In: OSDI, vol 8. pp 1–14
103. Moretti C, Bulosan J, Thain D, Flynn PJ (2008) All-pairs: an abstraction for data-intensive cloud computing. In: Parallel and distributed processing, 2008. IEEE international symposium on IPDPS 2008. IEEE, pp 1–11
104. Malewicz G, Austern MH, Bik AJC, Dehnert JC, Horn I, Leiser N, Czajkowski G (2010) Pregel: a system for large-scale graph processing. In: Proceedings of the 2010 ACM SIGMOD international conference on management of data. ACM, pp 135–146
105. Bu Y, Bill H, Balazinska M, Ernst MD (2010) Haloop: efficient iterative data processing on large clusters. Proc VLDB Endowment 3(1-2):285–296
106. Ekanayake J, Li H, Zhang B, Gunarathne T, Bae S-H, Qiu J, Fox G (2010) Twister: a runtime for iterative mapreduce. In Proceedings of the 19th ACM international symposium on high performance distributed computing. ACM, pp 810–818
107. Zaharia M, Chowdhury M, Das T, Dave A, Ma J, McCauley M, Franklin M, Shenker S, Stoica I (2012) Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In: Proceedings of the 9th USENIX conference on networked systems design and implementation. USENIX Association, pp 2–2
108. Bhatotia P, Wieder A, Rodrigues R, Acar UA, Pasquin R (2011) Incoop: mapreduce for incremental computations. In: Proceedings of the 2nd ACM symposium on cloud computing. ACM, p 7
109. Murray DG, Schwarzkopf M, Smowton C, Smith S, Madhavapeddy A, Hand S (2011) Ciel: a universal execution engine for distributed data-flow computing. In: Proceedings of the 8th USENIX conference on Networked systems design and implementation. p 9
110. Anderson TW (1958) An introduction to multivariate statistical analysis, vol 2. Wiley, New York
111. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Philip SY et al (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14(1):1–37
112. What analytics data mining, big data software you used in the past 12 months for a real project? (2012) http://www.kdnuggets.com/polls/2012/analytics-data-mining-big-data-software.html
113. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B (2008) KNIME: the Konstanz information miner. Springer
114. Sallam RL, Richardson J, Hagerty J, Hostmann B (2011) Magic quadrant for business intelligence platforms. CT, Gartner Group, Stamford
115. Beyond the PC. Special Report on Personal Technology (2011)
116. Goff SA, Vaughn M, McKay S, Lyons E, Stapleton AE, Gessler D, Matasci N, Wang L, Hanlon M, Lenards A et al (2011) The iplant collaborative: cyberinfrastructure for plant biology. Front Plant Sci 34(2):1–16. doi:10.3389/fpls.2011.00034
117. Baah GK, Gray A, Harrold MJ (2006) On-line anomaly detection of deployed software: a statistical machine learning approach. In: Proceedings of the 3rd international workshop on Software quality assurance. ACM, pp 70–77
118. Moeng M, Melhem R (2010) Applying statistical machine learning to multicore voltage & frequency scaling. In: Proceedings of the 7th ACM international conference on computing frontiers. ACM, pp 277–286
119. Gaber MM, Zaslavsky A, Krishnaswamy S (2005) Mining data streams: a review. ACM Sigmod Record 34(2):18–26
120. Verykios VS, Bertino E, Fovino IN, Provenza LP, Saygin Y, Theodoridis Y (2004) State-of-the-art in privacy preserving data mining. ACM Sigmod Record 33(1):50–57
121. van der Aalst W (2012) Process mining: overview and opportunities. ACM Transac Manag Inform Syst (TMIS) 3(2):7
122. Manning CD, Schütze H (1999) Foundations of statistical natural language processing, vol 999. MIT Press
123. Pal SK, Talwar V, Mitra P (2002) Web mining in soft computing framework, relevance, state of the art and future directions. IEEE Transac Neural Netw 13(5):1163–1177
124. Chakrabarti S (2000) Data mining for hypertext: a tutorial survey. ACM SIGKDD Explor Newsl 1(2):1–11
125. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30(1):107–117
126. Konopnicki D, Shmueli O (1995) W3qs: a query system for the world-wide web. In: VLDB, vol 95. pp 54–65
127. Chakrabarti S, Van den Berg M, Dom B (1999) Focused crawling: a new approach to topic-specific web resource discovery. Comput Netw 31(11):1623–1640
128. Ding D, Metze F, Rawat S, Schulam PF, Burger S, Younessian E, Bao L, Christel MG, Hauptmann A (2012) Beyond audio and video retrieval: towards multimedia summarization. In: Proceedings of the 2nd ACM international conference on multimedia retrieval. ACM, pp 2
129. Wang M, Ni B, Hua X-S, Chua T-S (2012) Assistive tagging: a survey of multimedia tagging with human-computer joint exploration. ACM Comput Surv (CSUR) 44(4):25
130. Lew MS, Sebe N, Djeraba C, Jain R (2006) Content-based multimedia information retrieval: state of the art and challenges. ACM Trans Multimed Comput Commun Appl (TOMCCAP) 2(1):1–19
131. Hu W, Xie N, Li L, Zeng X, Maybank S (2011) A survey on visual content-based video indexing and retrieval. IEEE Trans Syst Man Cybern Part C Appl Rev 41(6):797–819
132. Park Y-J, Chang K-N (2009) Individual and group behavior-based customer profile model for personalized product recommendation. Expert Syst Appl 36(2):1932–1939
133. Barragáns-Martínez AB, Costa-Montenegro E, Burguillo JC, Rey-López M, Mikic-Fonte FA, Peleteiro A (2010) A hybrid content-based and item-based collaborative filtering approach to recommend tv programs enhanced with singular value decomposition. Inf Sci 180(22):4290–4311
134. Naphade M, Smith JR, Tesic J, Chang S-F, Hsu W, Kennedy L, Hauptmann A, Curtis J (2006) Large-scale concept ontology for multimedia. IEEE Multimedia 13(3):86–91
135. Ma Z, Yang Y, Cai Y, Sebe N, Hauptmann AG (2012) Knowledge adaptation for ad hoc multimedia event detection with few exemplars. In: Proceedings of the 20th ACM international conference on multimedia. ACM, pp 469–478
136. Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci USA 102(46):16569
137. Watts DJ (2004) Six degrees: the science of a connected age. WW Norton & Company
138. Aggarwal CC (2011) An introduction to social network data analytics. Springer
139. Scellato S, Noulas A, Mascolo C (2011) Exploiting place features in link prediction on location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 1046–1054
140. Ninagawa A, Eguchi K (2010) Link prediction using probabilistic group models of network structure. In: Proceedings of the 2010 ACM symposium on applied Computing. ACM, pp 1115–1116

141. Dunlavy DM, Kolda TG, Acar E (2011) Temporal link prediction using matrix and tensor factorizations. ACM Transac Knowl Discov Data (TKDD) 5(2):10
142. Leskovec J, Lang KJ, Mahoney M (2010) Empirical comparison of algorithms for network community detection. In: Proceedings of the 19th international conference on World wide web. ACM, pp 631–640
143. Du N, Wu B, Pei X, Wang B, Xu L (2007) Community detection in large-scale social networks. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM, pp 16–25
144. Garg S, Gupta T, Carlsson N, Mahanti A (2009) Evolution of an online social aggregation network: an empirical study. In: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. ACM, pp 315–321
145. Allamanis M, Scellato S, Mascolo C (2012) Evolution of a location-based online social network: analysis and models. In: Proceedings of the 2012 ACM conference on Internet measurement conference. ACM, pp 145–158
146. Gong NZ, Xu W, Huang L, Mittal P, Stefanov E, Sekar V, Song D (2012) Evolution of social-attribute networks: measurements, modeling, and implications using google+. In: Proceedings of the 2012 ACM conference on Internet measurement conference. ACM, pp 131–144
147. Zheleva E, Sharara H, Getoor L (2009) Co-evolution of social and affiliation networks. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 1007–1016
148. Tang J, Sun J, Wang C, Yang Z (2009) Social influence analysis in large-scale networks. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 807–816
149. Li Y, Chen W, Wang Y, Zhang Z-L (2013) Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships. In Proceedings of the sixth ACM international conference on Web search and data mining. ACM, pp 657–666
150. Dai W, Chen Y, Xue G-R, Yang Q, Yu Y (2008) Translated learning: transfer learning across different feature spaces: In: Advances in neural information processing systems. pp 353–360
151. Cisco Visual Networking Index (2013) Global mobile data traffic forecast update, 2012–2017 http://www.cisco.com/en.US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html (Son erişim: 5 Mayıs 2013)
152. Rhee Y, Lee J (2009) On modeling a model of mobile community: designing user interfaces to support group interaction. Interactions 16(6):46–51
153. Han J, Lee J-G, Gonzalez H, Li X (2008) Mining massive rfid, trajectory, and traffic data sets. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, p 2
154. Garg MK, Kim D-J, Turaga DS, Prabhakaran B (2010) Multimodal analysis of body sensor network data streams for real-time healthcare. In: Proceedings of the international conference on multimedia information retrieval. ACM, pp 469–478
155. Park Y, Ghosh J (2012) A probabilistic imputation framework for predictive analysis using variably aggregated, multi-source healthcare data. In: Proceedings of the 2nd ACM SIGHIT international health informatics symposium. ACM, pp 445–454
156. Tasevski P (2011) Password attacks and generation strategies. Tartu University: Faculty of Mathematics and Computer Sciences