Chapter 2
Related Technologies

Abstract In order to gain a deep understanding of big data, this chapter will introduce several fundamental technologies that are closely related to big data, including cloud computing, Internet of Things (IoT), data center, and Hadoop. For each related technology, a general introduction is first provided highlighting their key features. Then the relationship between the technology and big data is examined in detail.

2.1 Cloud Computing

2.1.1 Cloud Computing Preliminaries

In the big data paradigm, reliable hardware infrastructures is critical to provide reliable storage. The hardware infrastructure includes masses of elastic shared Information and Communications Technology (ICT) resources. Such ICT resources shall be capable of horizontal and vertical expansion and contraction, and dynamic reconfiguration for different applications. Over the years, the advances of cloud computing have been changing the way people acquire and use hardware infrastructure and software services [1].

Cloud Computing is evolved from Distributed Computing, Parallel Computing, and Grid Computing, or a commercial realization of the computer-scientific concept. In a narrow sense, cloud computing means the delivery and use mode of IT infrastructure, i.e., acquiring necessary resources through the Internet on-demand or in an expandable way. In a general sense, cloud computing means the delivery and use mode of services, i.e., acquiring necessary services through the Internet on-demand or in an expandable way. Such service may related to software and the Internet, or others. In short, it refers to the case that users access a server through the network in a remote location and then use some services provided by the server.

M. Chen et al., *Big Data: Related Technologies, Challenges and Future Prospects*,
SpringerBriefs in Computer Science, DOI 10.1007/978-3-319-06245-7_2,

This concept mainly evolves from some mixed concepts such as virtualized public computing and infrastructure. The key components of cloud computing is illustrated in Fig. 2.1.

Services provided by cloud computing can be described by three service models and three deployment models. Such a combination has many important features, including self-service as required, wide network access, resource pool, rapidity, elasticity, and service management, thus meeting the requirements of many applications. Therefore, cloud computing will be instrumental for big data analysis and applications.

2.1.2 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. 2.1. Big data is the object of the computation operation and stresses the storage capacity and computing capacity of a cloud server. The main objective of cloud computing is to use huge computing resources and computing capacities under concentrated management, so as to provide applications with resource sharing at a granularity and provide big data applications with 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 allows effective management of big data; the parallel computing capacity by virtue of cloud computing can improve the efficiency of acquiring and analyzing big data.

Even though there are many overlapped concepts and technologies in cloud computing and big data, they differ in the following two major aspects. First, the concepts are different in the sense that cloud computing transforms the IT architecture while big data influences business decision-making, while 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 operates in the upper level supported by cloud computing and provides functions similar to those of database and efficient data processing capacity. As Kissinger, President of EMC, said, 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 virtualization technologies. Therefore, cloud

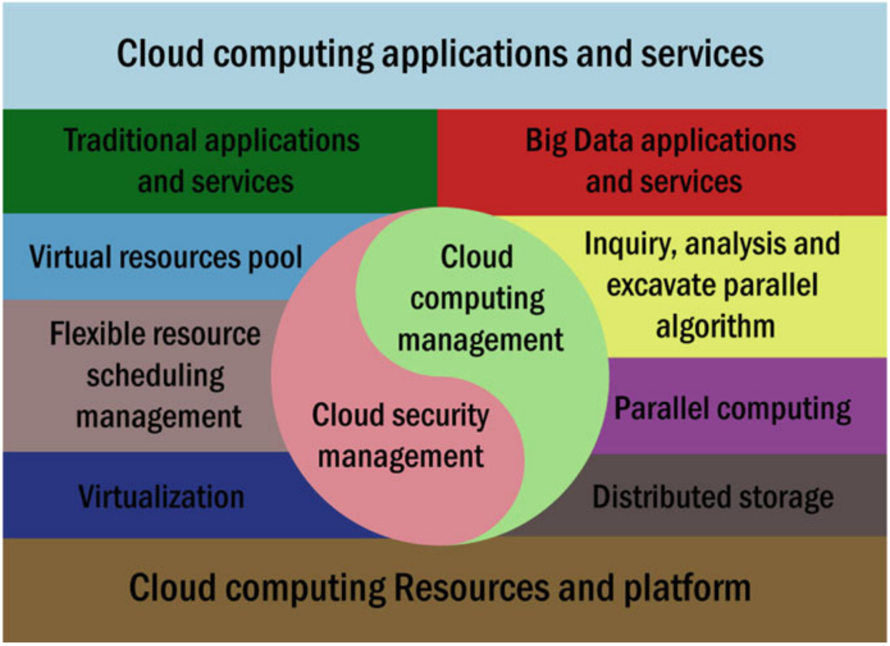
Cloud computing applications and services
Traditional applications and services
Big Data applications and services
Virtual resources pool
Cloud computing management
Inquiry, analysis and excavate parallel algorithm
Flexible resource scheduling management
Cloud security management
Parallel computing
Virtualization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

Fig. 2.1 Key components of cloud computing

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 IoT

2.2.1 IoT Preliminaries

The basic idea of IoT is to connect different objects in the real world, such as RFID, bar code readers, sensors, and mobile phones, etc., to realize information exchange and to make them cooperate with each other to complete a common task. The IoT architecture is illustrated in Fig. 2.2. IoT is deemed as the extension of the Internet and is an important part of the future Internet. IoT is mainly characterized with that it accesses every object in the physical world such that the objects can be addressed, controlled, and communicated with.

Compared with the Internet, IoT has the following main features [2].

- Various terminal equipments
- Automatic data acquisition
- Intelligent terminals

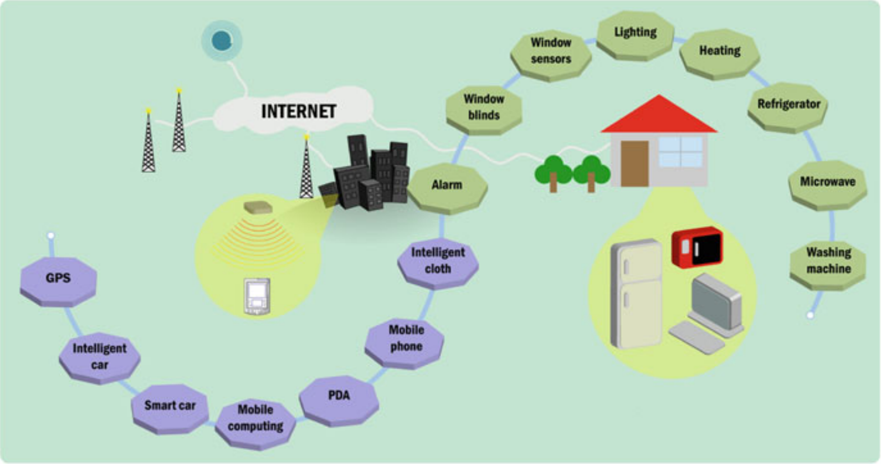

Fig. 2.2 Illustration of the IoT architecture

2.2.2 Relationship Between IoT and Big Data

In the IoT paradigm, an enormous amount of network sensors are embedded into devices 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 equipment in IoT.

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 rapid growth. 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 could 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: (a) abundant terminals generating masses of data; (b) data generated by IoT is usually semi-structured or unstructured; (c) 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 catch up with 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 IoT.

2.3 Data Center

In the big data paradigm, a data center is not only an organization 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 are mainly concerned with "data" other than "center." A data center 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 abundant development opportunities and great challenges to data centers.

- 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 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 data centers. In the big data paradigm, a data center shall not only be concerned 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.

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 in need [3].

2.4 Hadoop

2.4.1 Hadoop Preliminaries

Hadoop is a technology closely related to big data, which forms a powerful big data systematic solution through data storage, data processing, system management, and integration of other modules. Such technology has become indispensible to cope with the challenges of big data [4]. Hadoop is a set of large-scale software infrastructures for Internet applications similar to Google's FileSystem and MapReduce. Hadoop was developed by Nutch, an open-source project of Apache, with the initial design completed by Doug Cutting and Mike Cafarella. In 2006, Hadoop became an independent open-source project of Apache, which is widely deployed by Yahoo, Facebook, and other Internet enterprises. At present, the biggest Hadoop cluster operated by Yahoo has 4,000 sets of nodes used for data processing and analysis, including Yahoo's advertisements, financial data, and user logs.

Hadoop consists of two parts: HDFS (Hadoop Distributed File System) and MR framework (MapReduce Framework). HDFS is the data storage source of MR, which is a distributed file system running on commercial hardware and designed in reference to Google's DFS. HDFS is the basis for main data storage of Hadoop applications, which distributes files in data blocks of 64MB and stores such data blocks in different nodes of a cluster, so as to enable parallel computing for MR. An HDFS cluster includes a single NameNode for managing the metadata of the file system and DataNodes for storing actual data. A file is divided into one or multiple blocks and such blocks are stored in DataNodes. Copies of blocks are distributed to different DataNodes to prevent data loss. Apache HBase is a column-oriented storage, which imitates GooglesBigtable. Therefore, functions of Apache HBase are similar to those of BigTable as described in Part VI of HDFS. HBase may be taken as an input and output server of the MR task of Hadoop, and be accessed through Java API, REST, Avor, or Thrift APIs.

MR was developed similar to MapReduce of Google. The MR framework consists of one JobTracker node and multiple TaskTracker nodes. The JobTracker node is used for task distribution and task scheduling; TaskTracker nodes are used to receive Map or Reduce tasks distributed from JobTracker node and execute such tasks and feed task status back to the JobTracker node. MR framework and HDFS run in the same node set, so as to schedule tasks on nodes presented with data. Pig Latin is a high-level declarative language, which can describe the big data aggregation and analysis tasks in MR programming. Hive supports queries expressed by declarative similar to HiveQL and SQL. Hive introduces the concept of RDBMSs and SQL subset people are familiar with to Hadoop.

Apart from the aforementioned core parts, other modules related to Hadoop may also provide some supplementary functions required in the value chain of big data. Zookeeper and Chukwa are used to manage and monitor distributed applications run in Hadoop. It is worth noting that Zookeeper is the central service to maintain configuration and naming, provide distributed synchronization, and provide grouped services. Chukwa is responsible for monitoring system status and can display, monitor, and analyze collected data. Sqoop allows data to be conveniently passed between the structured data storage and Hadoop. Mahout is a data mining base executed on Hadoop using MapReduce. The base includes core algorithms of collaborative filtering used for clustering and sorting, and is based on batch processing.

Benefited from the huge success of the distributed file system of Google and the computational model of MapReduce for processing massive data, Hadoop, its clone, attracts more and more attentions. Hadoop is closely related to big data as nearly all leading enterprises of big data have commercial big data solutions based on Hadoop. Hadoop is becoming the corner stone of big data. Apache Hadoop is an open-source software framework. Hadoop realizes the distributed processing of massive data in the large-scale commercial server cluster, other than relying on expensive exclusive hardware and various systems to store and process data.

Hadoop has many advantages, but the following aspects are especially relevant to the management and analysis of big data:

- *Expandability*: Hadoop allows the expansion or shrinkage of hardware infrastructure without changing data format. The system will automatically re-distribute data and computing tasks will be adapted to hardware changes.
- *High Cost Efficiency*: Hadoop applies large-scale parallel computing to commercial servers, which greatly reduces the cost per TB required for storage capacity. The large-scale computing also enables it to accommodate the continually growing data volume.
- *Strong Flexibility*: Hadoop may handle many kinds of data from various sources. In addition, data from many sources can be synthesized in Hadoop for further analysis. Therefore, it can cope with many kinds of challenges brought by big data.
- *High Fault-Tolerance*: it is common that data loss and miscalculation occur during the analysis of big data, but Hadoop can recover data and correct computing errors caused by node failures or network congestion.

2.4.2 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 [5]. In addition, many companies provide Hadoop commercial execution and 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 [6] 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 bio-medicine to data-driven science. Gunarathne et al. in [7] 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: (a) assembly of genome segments; (b) 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 (i.e., MapReduce) framework may provide users an interface with more convenient services and reduce unnecessary costs.

References

1. Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, and Werner Vogels. Dynamo: amazon's highly available key-value store. In *SOSP*, volume 7, pages 205–220, 2007.
2. Luigi Atzori, Antonio Iera, and Giacomo Morabito. The internet of things: A survey. *Computer Networks*, 54(15):2787–2805, 2010.
3. Yantao Sun, Min Chen, Bin Liu, and Shiwen Mao. 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, 2013.
4. Tom White. *Hadoop: the definitive guide*. O'Reilly, 2012.
5. Wiki. Applications and organizations using hadoop. http://wiki.apache.org/hadoop/PoweredBy, 2013.
6. Arshdeep Bahga and Vijay K Madisetti. Analyzing massive machine maintenance data in a computing cloud. *Parallel and Distributed Systems, IEEE Transactions on*, 23(10):1831–1843, 2012.
7. Thilina Gunarathne, Tak-Lon Wu, Jong Youl Choi, Seung-Hee Bae, and Judy Qiu. Cloud computing paradigms for pleasingly parallel biomedical applications. *Concurrency and Computation: Practice and Experience*, 23(17):2338–2354, 2011.