Chapter 2 Big Data-Related Technologies

Big Data-Related Technologies

In order to gain a deep understanding of big data, this chapter will introduce several fundamental technologies that are closely related to Big data.

- Cloud computing
- Internet of Things (IoT)
- Data center and
- Hadoop

Cloud Computing

In the big data paradigm, reliable hardware infrastructures are critical to provide reliable storage.

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 the mode of services, i.e., acquiring necessary services through the Internet on-demand or in an expandable way. The key components of cloud computing are illustrated in the following Figure.

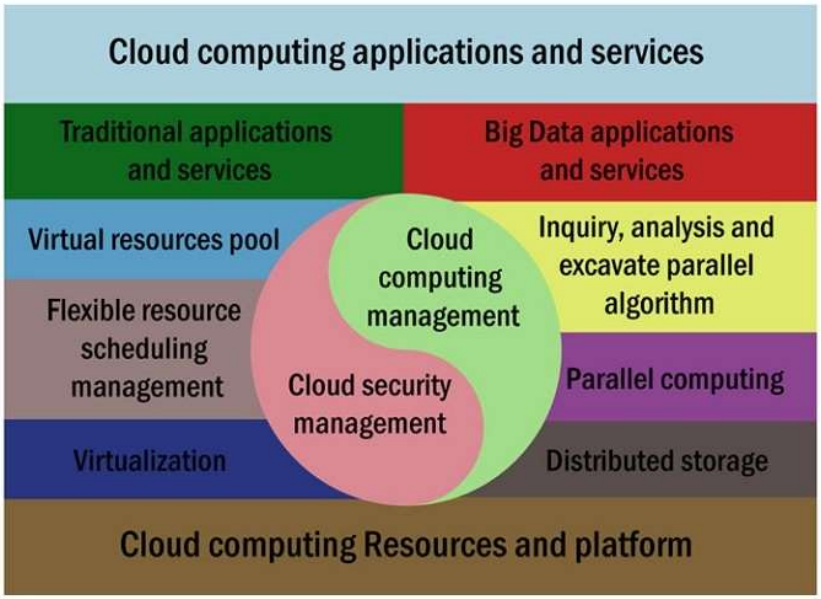

Figure: Key components of cloud computing

Big Data-Related Technologies

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.

Big Data-Related Technologies

Relationship Between Cloud Computing and Big Data

Cloud computing is closely related to big data. 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.

Big Data-Related Technologies

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.

Big Data-Related Technologies

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.

IoT

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.

RFID: Radio Frequency Identification

Big Data-Related Technologies

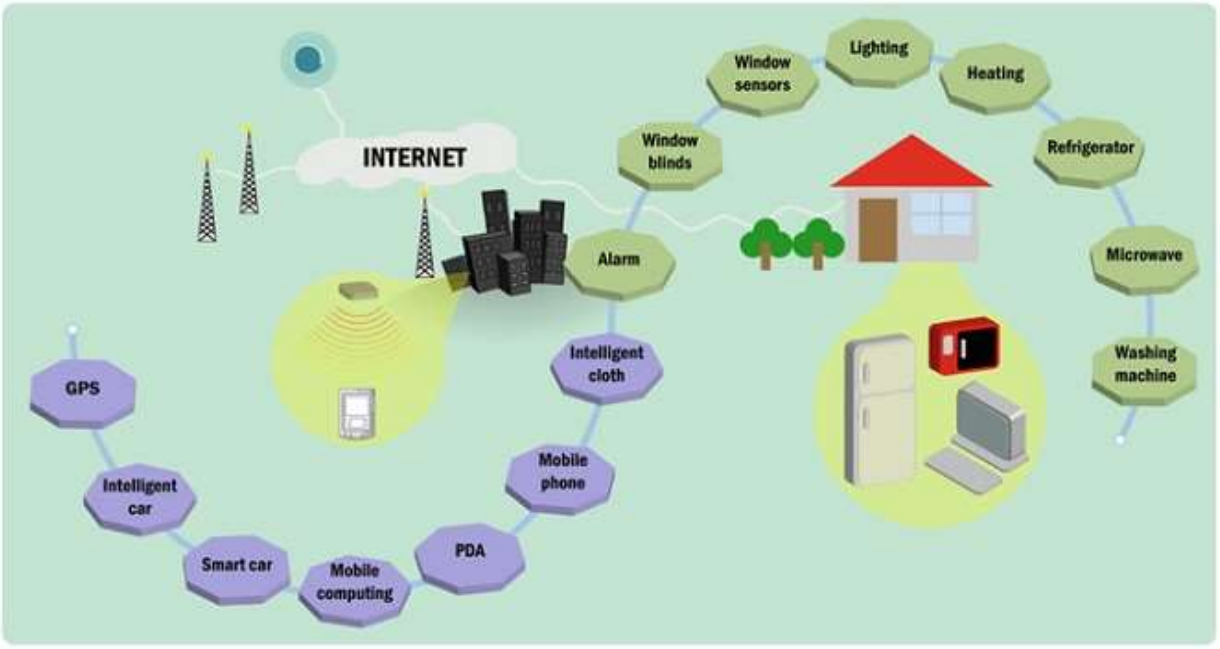

Figure: Illustration of the IoT architecture

Big Data-Related Technologies

IoT has the following main features - Various terminal equipments, Automatic data acquisition, Intelligent terminals.

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 equipment, 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

Big Data-Related Technologies

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.

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 (Report from Intel).

There is a compelling need to adopt big data for IoT applications, while the development of big data is already lagging behind. It has been widely recognized that these two technologies are interdependent and should be jointly developed.

Big Data-Related Technologies

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.

The emergence of big data brings about abundant development opportunities and great challenges to data centers.

Big data requires a data center to provide powerful backstage support. The big data paradigm has more stringent requirements on storage capacity and processing capacity, as well as network transmission capacity.

Big Data-Related Technologies

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 the data center.

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.

Big data is an emerging paradigm, which will promote the explosive growth of the infrastructure and related software of the data center.

Big Data-Related Technologies

Hadoop

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.

Hadoop is a set of largescale 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.

Big Data-Related Technologies

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.

Big Data-Related Technologies

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.

Big Data-Related Technologies

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 structured data storage and Hadoop. Mahout is a data mining base executed on Hadoop using MapReduce.

Big Data-Related Technologies

Hadoop Ecosystem

The Hadoop Ecosystem comprises:

- **HDFS** (Hadoop Distributed File System) which simply stores data files as close to the original format as possible.
- **HBase** is a Hadoop database management system and compares well with RDBMS. It supports structured data storage for large tables.
- **Hive** enables analysis of large data with a language similar to SQL, thus enabling SQL-type processing of data in a Hadoop cluster.
- **Pig** is an easy-to-understand data flow language, helpful in analyzing Hadoop- based data. Pig scripts are automatically converted to MapReduce jobs by the Pig Interpreter, thus enabling SQL-type processing of Hadoop data.
- **ZooKeeper** is a coordinator service for distributed applications.

Big Data-Related Technologies

The Hadoop Ecosystem comprises....

- **Oozie** is a workflow schedule system to manage Apache Hadoop Jobs.
- **Mahout** is a scalable machine learning and data mining library.
- **Chukwa** is a data collection system for managing large distributed systems.
- **Sqoop** is used to transfer bulk data between Hadoop and as structured data management systems such as relational databases.
- **Ambari** web-based tool for provisioning, managing and monitoring Apache Hadoop clusters.

Big Data-Related Technologies

Data Management	Data Access	Data Processing	Data Storage
Oozie (Workflow Monitoring)	Hive (SQL)	MapReduce (Cluster Management)	HDFS (Distributed File System)
Chukwa (Monitoring)	Pig (Data Flow)	Yarn (Cluster & Resource Management)	HBase (Column DB Storage)
Flume (Monitoring)	Mahout (Machine Learning)		
Zookeeper (Management)	Avio (RPC Serialization)		
	Sqoop (RDEMS Connector)		

Figure: Hadoop ecosystem elements at various stages of data processing

Big Data-Related Technologies

Advantage of Hadoop

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.

Big Data-Related Technologies

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.