# A Survey of Big Data for IoT in Cloud Computing

Junkuo Cao Mingcai Lin Xiaojin Ma

***Abstract*—Cloud computing provides a suitable platform for the application of big data in IoT. With the promotion of cloud platforms, big data in data intensive applications has rapidly developed in various fields and embodies unusual values on scientific research and economic benefits. Firstly, this paper introduces the concept of big data and cloud computing, and the main characteristics of big data are described in 6V mode. Subsequently, we focus on the main processes of big data, such as data acquisition, data organization and management, data analysis, and data presentation. Then, based on the architecture of could computing, some commonly used big data frameworks and their characteristics are described. Finally, application technologies of big data under cloud computing are introduced in some aspects, i.e., file system, data storage, data processing, analysis, and visualization. And the challenges for the future development of big data are discussed in the end.**

***Index Terms*—big data, cloud computing, data management, distributed computing**

## I. Introduction

With the continuous development of computer and network technology, a variety of massive data are daily generated from different fields, such as weather forecasting, e-commerce, health care, and other data intensive applications [1]. This is due not only to the close relationship between people's lives and information technology but also to the rapid progress of many new technologies, including the deployment of smart sensors [2][3], the development of the Internet of Things [4][5][6] and cloud computing [7]. From the individual point of view, these large volumes data that named as big data have different characteristics from traditional data sets, such as large amounts of data, diverse forms, high complexity, low value, and so on. But after data analysis, data mining, and other processing, we will get a lot of valuable information from these trivial data. However, data storage and processing based on traditional methods such as single-node or high performance cluster will not be able to process the big data effectively. Therefore, big data processing based on cloud computing has gradually become a popular trend.

With the rapid growth of big data, cloud computing as a new paradigm of computing has been more widely applied in various fields. Different from the traditional cluster system with fixed configuration, it uses virtualization technology to achieve unified management and maintenance of the underlying infrastructures and can provide users with ubiquitous, convenient and on-demand services that can be shared in a resource pool [8]. Cloud computing has three modes, such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). Based on the characteristics like virtualization, dynamic expansion, and heterogeneous distribution, users can obtain different high performance resources (such as storage, network, computing, etc.) on the cloud platform at a lower price according to their actual needs, while avoiding the huge investment in hardware infrastructures.

Due to the complexity and diversity of data types and huge scale of big data, it is difficult for traditional data storage and analysis systems to effectively solve this distributed and concurrency problem. At the same time, in the process of storage and analysis of big data, it is necessary to integrate advanced databases, high performance computing, and other technologies to improve the efficiency of data processing and data mining. With the help of cloud computing technology, it can not only provide flexible and diverse storage methods for massive data while improving the I/O speed, but also give full play to the powerful computing power of cloud platform and provide good technical support for distributed parallel computing. Moreover, with more scientific applications and life services migrating to the cloud platform, it will further promote the wide application of big data in various fields with low prices and efficient services.

Combined with the development of cloud computing, this paper introduces several big data technics and discusses the key issues of big data on cloud computing. The remainder of the paper is organized as follows. Section 2 introduces the concept and processing mechanism of big data. Section 3 presents the frame structure of big data in cloud computing. Then, the main technologies and challenges of big data are discussed in section 4. Finally, Section 5 concludes the paper.

## II. Big Data and Its Basic Process

The emergence of big data has changed people's views on traditional data and promoted the development of data acquisition, data storage, data analysis, and data mining. At present, big data has become a research hotspot in the field of information technology. Using big data can not only discover

Manuscript received October 22, 2019. This work was supported by Major Special Science and Technology Project of Hainan Province, China [Grant No. ZDKJ2017012].

Junkuo Cao is with Network and Data Center, Hainan Normal University, Haikou, China (e-mail: cjk@hainnu.edu.cn).

Mingcai Lin is with Haikou National Science Park, Hainan Normal University, Haikou, China (e-mail: 984279484@qq.com).

Xiaojin Ma is with School of Management, Henan University of Science and Technology, Luoyang, China. He is also with School of Computer Engineering and Science, Shanghai University, Shanghai, China (corresponding author, phone: (+86)13167020868, e-mail: xjma@shu.edu.cn).

information and patterns with high value but also promote the rapid development of related scientific research. Meanwhile, due to the complexity of big data and its wide application prospects, this field contains many universal and targeted research problems to be solved. Therefore, a brief introduction and the main processing of big data are presented as follows. The main characteristics and processes of big data are illustrated in Fig.1.

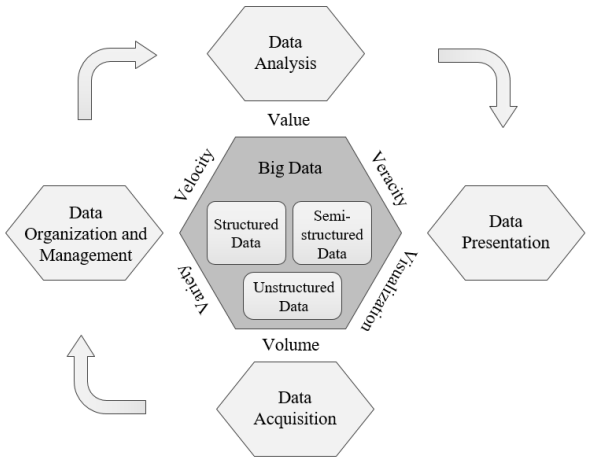


Fig.1 The main characteristics and processes of big data

*2.1 The Definition*

Having been around since the late 1980s, scientists have put forward the concept of big data and predicted that it will trigger data explosion in the information age [9] [10]. At present, both industry and academia are researching and developing big data. Researchers with various backgrounds have a different understanding of big data from different perspectives. This paper will focus on the efficient processing of big data in cloud computing from the abstract level, rather than on the study of a specific type of data. Therefore, big data in this paper is defined as large scale data sets composed of structured, unstructured and semi-structured data [11] [12].

Unlike traditional data, big data not only have a huge amount of data but also have more complex sources and structures. To better reflect the definition of big data, many scientists and experts define big data by the following main characteristics (called 3V and more Vs) [13] [14] [15].

• Volume (Data in the size): In daily production and life, many digital data are continuously generated from millions of devices and applications, such as sensors, cameras, IoT, financial transactions, scientific experiments, etc. According to relevant statistics, the amount of data generated in 2015 is about 8ZB, this value will probably be positive to 40zb by 2020 [16] [17].

• Variety (Data in the form): With the popularization of various information technologies, big data is no longer only has a traditional structured type, but also contains more complex semi-structured and unstructured data that are generated by logs, pictures, network text, sensors, and so on. While enriching data types, diversified data sources also lead to structural differences and semantic conflicts of data.

• Velocity (Data in the move): Data movement means not only the rapid generation or acquisition of data but also the efficient and rapid data processing. For many big data applications, a large amount of raw data will be entangled in a very short period, and the data need to be stored or processed in real-time to achieve the rapid flow of a series of processes from data acquisition, data storage, data retrieval, and data analysis.

• Value (Data in the refine): Despite the huge amount of data, the value density of big data is far from enough compared with traditional relational data. At the same time, the original data often cannot directly reflect the role and value but need further mining and analysis to extract effective information. This means that the value of big data lies in the correlation of data instead of each individual.

• Veracity (Data in the quality): In the process of raw data acquisition, some error information will be unavoidable. However, if big data only has the characteristics of large scale, diversity, and rapidity while lacking the necessary accuracy, it will greatly reduce the value of data. Only when the data have high accuracy, the actual value of big data can be reflected through its processing.

• Visualization (Data in the exhibition): How to display and communicate big data effectively in different environments has increasingly become an important research content. It has become the main way to display data content by using visualization methods such as charts, images, and animations. Visualization not only reflects the diversity of data forms but also reflects the changing trend and internal relationship between different data, which is easier to understand.

*2.2 Basic Process*

Through the analysis of definitions and characteristics, big data can be regarded as a systematic concept with a wide range of connotations. And big data needs periodic management when it is processed, usually including the processes of data acquisition, data organization and management, data analysis, and data presentation.

*2.2.1 Data Acquisition*

Data acquisition is not only the information source of different data systems but also the basis of subsequent data processing. Usually, data acquisition refers to the real-time or on-demand acquisition of data from different data sources into related systems. Due to the variety of types and rapid growth of big data, the multi-type characteristics of data need to be considered comprehensively in data acquisition. Meanwhile, effective technologies should be adopted to ensure the reliability and validity of the collected data to avoid the impact on subsequent processing. However, in the big data environment, it is difficult to avoid some errors and inaccuracies in the data acquired in different ways. Therefore, data cleaning (data preprocessing) is needed to ensure the quality of data.

*2.2.2 Data Organization and Management*

After preprocessing, the data from various application scenarios can effectively eliminate invalid data caused by noise, conflict, missing, and other factors. However, the amount of preprocessed data is still so large that traditional

storage and management technologies cannot meet the needs. Therefore, the management system of the distributed environment is used to achieve efficient access to big data. Besides, when large scale application data are stored in distributed nodes, data migration will bring huge challenges to bandwidth and network latency. To ensure the integrity and validity of data, the backup of data will also have a higher cost for storage and management. Therefore, in large scale distributed systems, the unified distributed storage and management of massive data is an important issue in big data applications.

*2.2.3 Data Analysis*

Preprocessed data are usually stored in distributed systems. Usually, more effective information cannot be obtained directly from these discrete stored data, which also reflects the low value characteristics of big data. More valuable information can be found through the distributed analysis, processing, deep mining and learning of the existing massive data. Therefore, data analysis is a key link in big data processing. Its processing technologies are not only related to the storage and management of data but also affected by the actual application types. Common data analysis includes machine learning, data mining, artificial intelligence, and other statistical analysis techniques.

*2.2.4 Data Presentation*

After analyzing or calculating data, how to display the results intuitively and effectively is another important issue. Data presentation mainly displays the data process, calculation results, and statistical trends scientifically in the form of text, voice, image, and other ways. At present, visualization and interaction are usually adopted to reflect the processing results of big data. Big data applications require visualization technologies based on their characteristics to quickly collect, screen, summarize and display the information needed by decision-makers. Meanwhile, the information displayed usually needs to be updated in real-time according to the latest data.

*2.3 Big Data Applications*

After processing much raw data, big data produce conclusions for reference and prediction after data analysis. Then users can use the existing data to predict future problems or trends, that will help them make decisions. Here are some common scenarios of big data applications:

Government management: The government usually collects and analyses the historical data to forecast the future development trend that can help to develop appropriate policies. Big data can effectively help government agencies formulate reasonable and feasible implementation plans, such as using the census to improve living conditions, using hydraulic data to prevent natural disasters, using food traceability management to track food safety, and so on.

Education industry: The education sector needs to continuously invest in infrastructure construction and improve the form of education to provide a good education for students. Big data can help educational institutions make use of these data to make up for the weaknesses in the teaching process. They use data to achieve better development of education. It also helps students to find their shortcomings and improve learning efficiency.

Medical care: Health departments and researchers make statistical analysis of various data, such as drug testing, patient records, and treatment plans. The results will help improve treatment and provide better medical assistance to patients. It can also improve the efficiency of administration, cost management, resource allocation and supply management in medical institutions.

Manufacturing: Conventional and sustained production in manufacturing will generate a lot of data. Big data applications can help producers reduce costs and waste, and help them produce high quality products in a shorter time. Meanwhile, the analysis of big data can enable enterprises to produce based on future demand forecasting to achieve higher profits.

Retail industry: The retail industry is an industry directly related to customers. It records many data from customers, including customer behavior, preferences, and biological data. Usually, it is difficult to obtain valuable information from the data of one or a few customers. However, many conclusions that will help improve sales can be gained through the big data. For example, business development can be targeted according to the analysis of sales behavior data in a certain area. Big data provides accurate information to solve complex problems and plays an important role in current business activities.

## III. Big Data on Cloud Computing

With the explosive growth and various characteristics of big data, traditional data processing technology has been difficult to effectively solve the data processing problems in various fields. Therefore, various big data processing systems have been proposed. From different processing forms, these processing systems can be divided into batch data processing systems, stream data processing systems, interactive data processing systems, and graph data processing systems [18]. These different big data processing systems, which help users efficiently carry out data processing, data mining, and other tasks, often need a certain scale of hardware infrastructure as support. Especially for individuals or minor enterprises, hardware platforms that can meet the needs of big data processing systems are often expensive to invest, and there are high barriers to the optimization of deployment platforms. The emergence of cloud computing provides an effective and feasible solution to these difficulties in big data processing.

*3.1 Cloud Computing*

As a new model, cloud computing has become a mainstream resource service platform. It provides a convenient and fast way to access different resources (such as storage, network, computing, application, etc.) on-demand from the configurable resource pool [19]. Fig.2 shows a brief system framework of cloud computing.

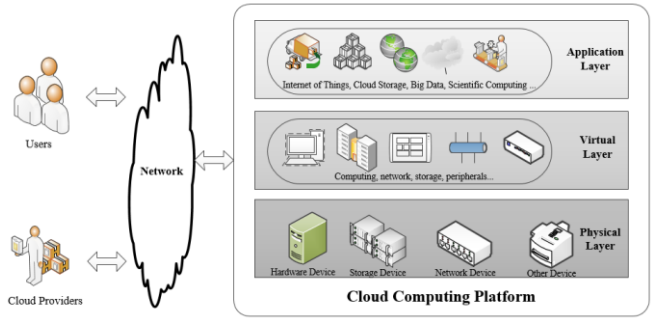


Fig.2 Cloud computing framework

As shown in Fig.2, the main framework of cloud computing can be divided into three parts: the physical layer, the virtual layer, and the application layer. First, the physical layer refers to the underlying infrastructure of cloud computing, which is the hardware support and network deployment unit of the entire cloud platform. The physical layer covers many hardware resources such as storage, network, and servers, providing powerful computing and storage capabilities for cloud computing platforms. Second is the virtual layer, which uses virtualization technology to allocate and deploy all kinds of physical resources. Virtualization technology is the core of the cloud computing framework. It redefines and divides hardware resources through software. And the entities like servers, networks, storage in cloud computing are abstracted and transformed in the way of the virtual resource pool, thus realizing dynamic allocation, flexible scheduling, and cross-domain sharing of resources. The last is the application layer, including user interfaces, big data applications, resource management systems, fault detection systems, and other applications. The application layer provides specific platforms for different domains. Users can deploy applications on cloud according to their needs, and realize development, computing, storage, and other services in different environments.

### *3.2 Frame Structure*

Big data and cloud computing can be regarded as a complementary and inseparable relationship. Cloud computing should be used as a supporting platform for big data processing. Meanwhile, the value of big data can make cloud computing better integrate with different industrial applications. Furthermore, cloud computing provides storage and computing resources to support the processing of big data. Through data query and data analysis, big data produces necessary value information for related applications in the cloud.

Fig.3 illustrates the processing and analysis of big data in cloud computing. Users upload all kinds of data to the data center of the cloud through the web or application side and store them in a distributed file system. Subsequently, users use cluster-based parallel processing algorithms to process data at high speed and present the results in different forms like texts, voices, and charts. In this process, due to the complexity of types and the huge amount of data, it is necessary to rely on some technologies like distributed storage, distributed processing and virtualization to achieve effective processing.

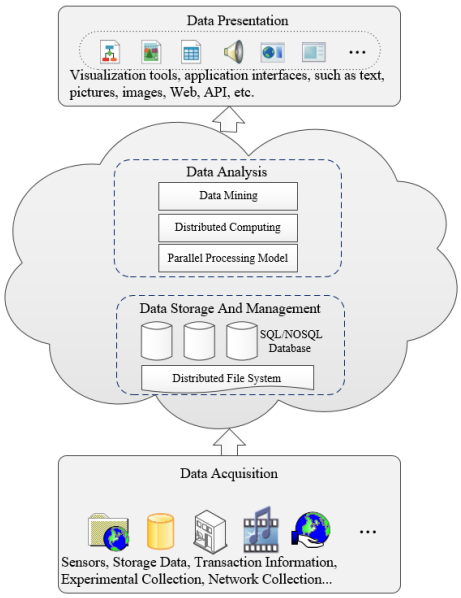


Fig.3 Big data processing architecture based on cloud

In the distributed environment, the processing needs a unified framework to deal with the allocation and coordination of storage, computing, load balancing, data migration, and other tasks. Fig.4 shows some common big data frameworks in different processing modes. The briefly introduces are as follows.

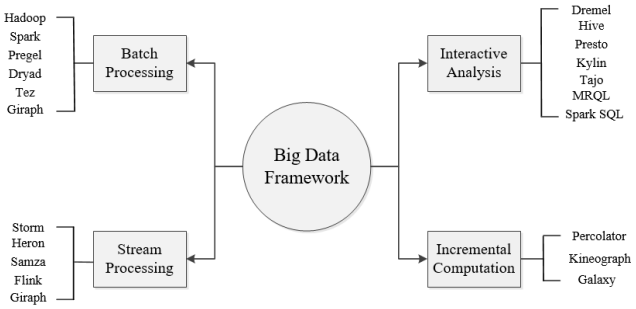


Fig.4 Common big data frameworks

#### *3.2.1 Hadoop*

Hadoop is a big data processing framework dedicated to batch processing. It mainly includes three cores: distributed file system (Hadoop distributed file system, HDFS), resource scheduling (yet another resource negotiator, YARN), and computing framework (MapReduce) [20]. HDFS is a distributed file system that coordinates storage and replication among cluster nodes. HDFS guarantees high fault tolerance of data by maintaining multiple copies of working data, and it can provide high throughput of data access. Yarn is a general resource management system, which can provide unified resource management and scheduling for upper application. This component is responsible for managing the underlying resources, scheduling jobs, improving cluster performance, sharing data, and so on. MapReduce is a computing model and method for parallel processing of big data. It combines a bunch of disorderly data according to some characteristics and then gets the result after processing.

*3.2.2 Spark*

Spark is a batch processing framework with stream processing capabilities [21]. The basic principle of Spark is to divide streaming data into small time segments and process them in a batch-like manner. And its advanced application interface strips away the focus on the cluster itself while supporting interactive computing and complex algorithms. Compared with Hadoop, Spark has faster processing speed through memory computing. Also, Spark supports the use of directed acyclic graphs to develop complex multi-step data pipelines, and the shared memory data across directed acyclic graphs enable different jobs to process the same data together. Furthermore, Spark is a general engine, which can be used to solve many important tasks and problems in the field of big data, including offline batch processing, interactive query, real-time flow computing, machine learning, graph computing, and so on.

*3.2.3 Storm*

Storm is a real-time data processing framework with low latency, high availability, scalability, data loss, and other characteristics [22]. Its core components include nimbus, supervisor, worker, and task. Nimbus is mainly responsible for resource allocation and task scheduling. A supervisor is responsible for receiving tasks allocated by nimbus and starting/stopping its worker process. The worker runs specific processing logic processes, which can be divided into spout and blot. And each spout/blot becomes a task thread. The Storm framework ensures that every message is processed, and it is suitable for flow processing workloads with high latency requirements.

*3.2.4 Samza*

Samza is a flow processing framework. It uses open source distributed message processing system apache Kafka to implement message service and uses its manager yarn to implement fault-tolerant processing, processor isolation, security, and resource management [23]. Samza is suitable for real-time streaming data processing services, such as data tracking, log services, real-time services, and other applications. In the process of stream data processing, each Kafka cluster is connected to a cluster that can run yarn and cares for Samza operations. The main features of Samza include providing a simple callback-based message processing interface, realizing stateful message management by storing historical data, ensuring orderly processing and persistence of messages, partitioning and distributing each layer structure, and so on.

*3.2.5 Flink*

The core of Flink is a streaming data stream execution engine, which provides data distribution, data communication, and fault-tolerant mechanism for distributed computing of data stream [24]. Designed to run in a common cluster environment, Flink can perform computation at memory speed and any scale. Client, job manager, and task manager are Flink's basic components, in which the client submits tasks to the job manager, and the job manager distributes tasks to different task managers for performing. Flink can handle unbounded and bounded data sets, and even provide accurate results in the case of disorderly or delayed loading. It can deploy applications everywhere with good throughput and low latency when thousands of nodes run on a large scale.

## IV. Technologies and Challenges of Big Data in Cloud

The demand for big data in various fields mainly concentrates on two aspects: real-time processing of massive data and generating effective information from data mining. The diversity of massive data in a distributed environment makes data processing have new characteristics, and its technical requirements are quite different from traditional systems. Based on data acquisition, storage management, processing analysis, and result presentation, some cloud-based technologies have been developed to deal with big data in the cloud. Meanwhile, with the development of various new technologies and the promotion of big data applications, some challenges also appear in the process of big data processing in cloud computing.

*4.1 Technologies*

*4.1.1 Distributed File System*

The distributed file system refers to the multiple resources that make up the file system that can be distributed in different nodes and communication between nodes through the network. The design of a distributed file system is based on client/server mode, which is due to the data scale effect. The data volume of big data has far exceeded the storage capacity of a local single machine, so it needs to be stored and managed separately. When using a distributed file system, users need not pay attention to which node the data are stored on or obtained from, but manage and store all kinds of data just like using a local file system.

HDFS is a commonly used distributed file system, which is highly fault-tolerant, suitable for deployment on inexpensive machines and can provide high throughput [25]. HDFS adopts a master-slave structure model, and an HDFS cluster is composed of a namenode and several datanodes. The namenode serves as the primary server to manage the namespace of the file system and the access operations of the client to the files, while datanode in the cluster manages the stored data. Furthermore, HDFS store files in block sequence and distribute them evenly in clusters after copying to ensure the high availability of data.

*4.1.2 Scalable Data Storage*

The mass data from different sources need to be stored and managed effectively in the cloud. Since these data are usually unstructured or semistructured with different scales, traditional relational databases are not suitable for them. Unlike relational databases, NoSQL databases allow loose structure and highly flexible data models and are increasingly used in big data and real-time network applications. NoSQL databases do not guarantee the ACID characteristics of relational data, and it has characteristics like easy expansion, large data volume, high performance, simple structure, and so on [26]. According to different types, NoSQL databases can be divided into the following categories:

Key-Value Storage Database: This type of database mainly uses a hash table with a specific key and a pointer pointing to

specific data. The advantage of the Key/Value model is the fast search speed and deployment convenience. Such databases include Tokyo Cabinet, Redis, Voldemort, Oracle BDB.

Column Storage Database: Such databases are usually used to store distributed mass data. The characteristic of these databases is that keys point to multiple columns, which are arranged by column families. Cassandra, HBase, and Riak are some common column storage databases.

Document Database: Document database can be regarded as upgraded versions of key-value storage databases, allowing nesting of key values. This type of data model is versioned documents, and semi-structured documents are stored in specific formats like JSON. CouchDB and MongoDB are two document databases.

Graph database: Graphical database is different from rigid SQL databases and other types of databases. It uses a flexible graphical model and can extend to multiple servers. However, due to the need to compute the whole graph structure, it is not suitable for distributed cluster applications.

*4.1.3 Data Processing*

The processing of big data based on the cloud can be divided into real-time stream processing and large-scale batch processing. Each method has its advantages and disadvantages, but how to choose it depends on the specific business need. Real-time stream processing is used to process continuous data in real-time, which is the key to convert big data into fast data. The data to be processed are continuously arriving and need to be processed quickly without delay. For example, a detector is set up somewhere in a river to detect the situation of water (data), which is a real-time stream processing application. And large-scale batch processing is the processing of data sets that are grouped in a specific time interval. Usually, the data processed by batch processing need to be generated in a certain period and loaded into a certain type of storage, database or file system before processing.

As mentioned in section 3, Hadoop is a batch processing framework, Storm and Samza are stream processing framework, while Spark and Flink are hybrid frameworks capable of batch and stream processing. The batch functionality of Hadoop comes from the MapReduce engine, whose processing technology adopts the way of key-value pairs. The basic process is: data set is divided into small blocks and allocated to all available nodes after reading from the file system, then the result is combined by grouping the intermediate computing state at each node. In contrast, real-time stream processing systems can handle almost unlimited amounts of data. But at the same time, only one or very few (micro-batch processing) data can be processed, and just the minimum state can be maintained between different records.

*4.1.4 Analytics and Visualization*

Data analysis generally means using appropriate statistical analysis methods or tools to analyze collected data and converts it into information about historical patterns, current performance or future trends. Unlike the traditional data analysis, the analysis based on big data refers to the use of new processing mode to extract useful information and form conclusions within the affordable time range. For big data analysis, the most important challenge is to extract useful knowledge from large amounts of existing data. This approach usually involves data mining, including association rule learning, cluster analysis, classification, and regression. Besides, machine learning is another fundamental technology in big data analysis. Machine learning uses computers to simulate human learning and uses existing knowledge data to improve learning efficiency. The goal is to automatically identify complex patterns based on data and make informed decisions. Machine learning is highly correlated with data mining, usually using the same technical methods, such as classification, clustering, and regression. However, data mining focuses on discovering unknown attributes of data from big data, while machine learning focuses on predicting from training data.

Visualization is an important component of big data statistical analysis. Effective visualization forms such as images, charts, and animations can help users analyze data effectively and accurately. In order to meet the actual needs, big data visualization applications should generally have the following characteristics: can handle different types of incoming data, can adjust the results through different filters, can interact with the data set in the analysis process, can provide user collaborative options. Some big data visualization platforms include Tableau, Orange, Power BI, IBM Watson Analytics, and SAP Leonardo.

*4.2 Challenges*

Big data has been rapidly promoted and widely used depending on the convenience of cloud computing, but it also faces various challenges. Big data based on cloud computing has the characteristics of mass and diversity, which makes the way of data collection, data management, and data analysis in distributed environments different from traditional systems. Its data processing and scalability bring new requirements to related technologies.

*4.2.1 Data Collection*

In the actual production environment of various fields, data is usually stored in separate data warehouses of different departments, which lead to data fragmentation. In the processing of data collection, it is necessary to associate and integrate different types of data, while sharing technologies and tools, to eliminate the phenomenon of information island as far as possible. Meanwhile, due to the diversity of original data formats and different sources, data preprocessing technology needs to be further improved to avoid the irregularity and unreasonably of data, while will improve the quality, accuracy, and availability of the collected data.

*4.2.1 Data Management*

Depending on the rapid development of cloud computing, big data is managed by distributed file systems. To improve the efficiency of parallel processing, resource sharing, communication mechanism, query optimization, and other technical schemes among nodes still need further study. For mass unstructured data and semi-structured data, the current key-value pair method can meet different types of data storage needs, but its structure is not easy to perform

connection, query, modification, and other related operations between tables. Furthermore, based on data distributed storage, it is significant to improve the operation methods (such as aggregation, statistics, sorting, etc) in MapReduce and other models. For example, the current key/value-based indexing technology relies on business keywords, and non-key queries still need full table scanning that reduces efficiency. Meanwhile, there is no unified query optimization algorithm to solve the query requirements under complex conditions.

#### *4.2.2 Data Analytics*

The core of big data analysis is a data mining algorithm, which based on different data types and formats that will present different characteristics of data. Considering the characteristics of big data, three key points need to be considered in data analysis. Firstly, although the original data has been preprocessed, there are still some imperfections in different types of data. So, it is necessary to minimize the impact of the inaccuracy of data in the processing of data analysis. Secondly, generation time and volume of data are very difficult to predict and calculate, and a dynamic response mechanism is needed to solve this uncertainty problem. Especially in the cloud, data storage and computing resources should be reduced as much as possible through reasonable scheduling and configuration to achieve the goal of obtaining ideal analysis results at the minimum cost. Finally, the existing big data analysis tools can only focus on a few aspects. A general analysis software based on the characteristics of massive, multi-structure and real-time data will help to promote the development of big data.

#### *4.2.3 Data Security*

After realizing the distributed storage of big data in the cloud, how to ensure the integrity, availability, and confidentiality of network data while avoiding information leakage has become the core issue to be considered in various fields. On the one hand, the security of big data lies in the strict management of daily operations. For example, desensitization of data is very necessary for the process of data collection and storage, while sensitive data should be classified and managed according to the actual situation. Also, detailed logs should be recorded in all data operations, and uniform control of access control with fine-grained authorization should be realized. On the other hand, vulnerability scanning and risk prevention of big data development software need to be carried out regularly. Security of big data needs to cover every aspect of the implementation process, whether hardware, platform, software or data needs to achieve comprehensive integrated security protection. Besides, many new technologies have been rapidly developed and applied to the demand for big data processing.

## V. Conclusion

With the rapid development of network technologies and data intensive applications, massive data information is produced every day. And how to make full use of these large-scale, multi-mode data, mining its value, has become an important research content of big data.

According to the basic concepts and main characteristics of big data and cloud computing, this paper introduces the definition of big data and gives a brief description of the main processing process. The development of big data is inseparable from the cloud platform, and big data is also an important application in cloud computing. Different from traditional computing modes, cloud computing has become the basic platform for big data applications with the characteristics of on-demand acquisition, virtualization, and scalability. A distributed file system based on cloud realizes the storage and management of massive data. Unlike traditional relational data, unstructured and semi-structured data are managed and maintained in different frameworks by the model of NoSQL.

## References


[1] S. J. Walker, “Big Data: A Revolution That Will Transform How We Live, Work, and Think,” International Journal of Advertising, no. 33:1, pp. 181-183, 2014.

[2] T. Zhu, Q. Ma, S. Zhang and Y. Liu, “Context-free Attacks Using Keyboard Acoustic Emanations”, in Proceedings of ACM Conference on Computer and Communications Security, pp. 453-464, 2014.

[3] B. Sun, Q. Ma, S. Zhang, et al., “iSelf: Towards Cold-Start Emotion Labeling Using Transfer Learning with Smartphones”, ACM Transactions on Sensor Networks, Volume 13 Issue 4, Article No. 30, December 2017.

[4] J. Gubbi, R. Buyya, S. Marusic, et al., “Internet of Things (IoT) A vision, architectural elements, and future directions,” Future Generation Computer Systems, vol. 29, no. 7, pp. 1645–1660, 2013.

[5] Q. Ma, K. Liu, X. Xiao and et al., “Link Scanner: Faulty Link Detection for Wireless Sensor Networks”, in Proceedings of IEEE INFOCOM, 2013.

[6] K. Liu, Q. Ma and Y. Liu, Self-diagnosis for large wireless sensor networks, in Proceedings of IEEE INFOCOM, pp. 5535-5545, 2011.

[7] X. Zhu, J. Wang, H. Guo, et al., “Fault-Tolerant Scheduling for Real-Time Scientific Workflows with Elastic Resource Provisioning in Virtualized Clouds,” IEEE Transactions on Parallel and Distributed Systems, vol. 27, no. 12, pp. 3501-3517, 2016.

[8] P. Mell, and T. Grance, “The NIST Definition of Cloud Computing,” Communications of the Acm, vol. 53, pp. 50-50, Jun, 2010.

[9] A. Toffler, The Third Wave: Bantam, 1984.

[10] M. Cox, and D. Ellsworth, “Application-Controlled Demand Paging for Out-of-Core Visualization,” in Proceedings. Visualization '97 (Cat. No. 97CB36155), Phoenix, AZ, USA, USA, 1997.

[11] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, et al., “Big Data technologies A survey,” Journal of King Saud University – Computer and Information Sciences, vol. 30, pp. 431-448, 2018.

[12] K. He, L.-J. Zhang, S. Dustdar, et al., “Guest Editorial Cloud Services Meet Big Data,” IEEE Transactions on Services Computing, vol. 11, no. 4, pp. 599-601, 2018.

[13] B. Furht, and F. Villanustre, "Introduction to Big Data," Big Data Technologies and Applications, Springer International Publishing, 2016, pp. 3-11.

[14] A. Mohamed, M. K. Najafabadi, Y. B. Wah, et al., “The state of the art and taxonomy of big data analytics: view from new big data frame-work,” Artificial Intelligence Review, February 01, 2019.

[15] C. K. Emani, N. Cullot, and C. Nicolle, “Understandable Big Data: A survey,” Computer Science Review, vol. 17, pp. 70-81, 2015.

[16] V. Rajaraman, “Big data analytics,” Resonance, vol. 21, no. 8, pp. 695-716, August 01, 2016.

[17] R. Kune, P. K. Konugurthi, et al., “The anatomy of big data computing,” Software: Practice and Experience, vol. 46, no. 1, pp. 79-105, 2016.

[18] C. XQ, J. XL, W. YZ, et al., “Survey on Big Data System and Analytic Technology,” Journal of Software, vol. 25, no. 9, pp. 1240-1252, 2014.

[19] Y. Wang, J. Li, and H. H. Wang, “Cluster and cloud computing framework for scientific metrology in flow control,” Cluster Computing, vol. 22, no. 1, pp. 1189-1198, January 01, 2019.

[20] D. Bhattacharya, F. Currim, and S. Ram, “Evaluating Distributed Computing Infrastructures: An Empirical Study Comparing Hadoop Deployments on Cloud and Local Systems,” IEEE Transactions on Cloud Computing, vol. PP, no. 99, pp. 1-14, 2019.

[21] M. Zaharia, M. Chowdhury, M. J. Franklin, et al., “Spark: cluster computing with working sets,” in Proceedings of the 2nd USENIX conference on Hot topics in cloud computing, Boston, MA, 2010, pp. 10-10.
[22] P. Karunaratne, S. Karunasekera, and A. Harwood, “Distributed stream clustering using micro-clusters on Apache Storm,” Journal of Parallel and Distributed Computing, vol. 108, pp. 74-84, 2017.
[23] M. Kleppmann, "Samza," Encyclopedia of Big Data Technologies, S. Sakr and A. Zomaya, eds., Springer International Publishing, 2018, pp. 1-8.
[24] D. García-Gil, S. Ramírez-Gallego, S. García, et al., “A comparison on scalability for batch big data processing on Apache Spark and Apache Flink,” Big Data Analytics, vol. 2, no. 1, pp. 1, March 01, 2017.
[25] D. Huang, D. Han, J. Wang, et al., “Achieving Load Balance for Parallel Data Access on Distributed File Systems,” IEEE Transactions on Computers, vol. 67, no. 3, pp. 388-402, 2018.
[26] A. Colaso, P. Prieto, J. et al., “Memory Hierarchy Characterization of NoSQL Applications through Full-System Simulation,” IEEE Transactions on Parallel and Distributed Systems, vol. 29, no. 5, pp. 1161-1173, 2018.

**Junkuo Cao** received the Ph.D. degree from the Department of Computer Science and Engineering, Fudan University, in 2009. He is currently an Associate Professor in Network and Data Center, Hainan Normal University. His research interests include intelligent information processing and Big Data Mining, with a special focus on natural language processing and Chinese information processing.

**Mingcai Lin** received his bachelor's degree in 1996 and master's degree in 2010 from University of Science and Technology Beijing. He is currently the Director of Haikou National Science Park, and an Associate Professor in the School of Information Science and Technology, Hainan Normal University. The objective of his research is Data Mining in Small and medium-sized enterprises (SME) Services.

**Xiaojin Ma** received the BS, MS in Computer Science and Management Science & Engineering from Henan University of Science and Technology in 2003, 2013, respectively. He is working toward the Ph.D. degree in Shanghai University, China.