ISSN 2616-5805 Vol. 4, Issue 5: 54-61, DOI: 10.25236/AJMS.2023.040508

Research on Big Data Acquisition Method Based on Mapreduce Algorithm

Jiaxi Guo[1,a,*], Chuansheng Wu[1,b]

[1]University of Science and Technology Liaoning, Anshan, China
[a]3414324108@qq.com, [b]gykwcs@163.com
**Corresponding author*

Abstract: *With the progress of science and technology, big data has become an indispensable part of our lives. The application fields of big data are very wide, from business decision-making, medical research to education, all of them are involved. In this paper, we will focus on the acquisition of big data, methods and algorithms. A big data analysis method, a system, a computer device, and a storage medium thereof, relate to the technical field of big data analysis. The method obtains the big data information accessed by the user during the target time period; and stores the big data information in a distributed database in slices according to time; then performs data analysis tasks based on predetermined rules on the big data information stored in the distributed database, and obtains the analysis results; configures a data caching server to cache the analysis results; and when obtaining a data query request from the front-end, the parameters of the request are cached in the data caching server according to the parameters of the data query request. When the front-end data query request is obtained, the corresponding analysis result data is called in the data cache server as the query result, and the query result is output to the front-end and visualized, which can reduce the burden of the processor of the big data analysis system, improve the access speed of the user, avoid the access lagging, and ensure the smoothness of the user's access.*

Keywords: *data analysis methods, systems, computer equipment, storage media*

1. Preamble

There are many ways to obtain big data, mainly relying on data collection tools and technologies. First, we can use traditional data collection methods, such as questionnaires and interviews. However, with the growth of data volume, these methods have been unable to meet the demand. Therefore, we need to use more efficient data collection tools and technologies, such as sensors, remote sensing technology, etc. In addition, with the development of the Internet, tools such as web crawlers and APIs also provide us with new ways to obtain big data. No matter what method is used, attention should be paid to the accuracy and integrity of data when obtaining big data[1].

With the advent of the era of big data, the amount of information in the network is growing exponentially, which brings about the problem of information overload; Recommendation system is one of the most effective ways to solve information overload. Big data recommendation system has gradually become a research hotspot in the information field, and the "big data era" has come. With the arrival of the "big data" era, people are mining and using massive data, which indicates the arrival of a new wave of productivity growth and consumer surplus. Big data is another disruptive technological revolution in the IT industry after cloud computing and the Internet of Things. With the advent of big data, people's demand for big data continues to increase, which will increase the burden of the intelligent analysis system and cause the burden of the intelligent analysis system processor. When the intelligent analysis system has a burden, it will reduce the user's access speed, cause a jam, or fail to load the data, which will lead to poor access for users[2].

2. Big data analytics

After acquiring big data, we need to adopt appropriate methods to process and analyze the data. The processing and analysis of big data need to be combined with specific application scenarios and purposes. In business decision-making, data analysis can help enterprises understand market trends and optimize products and services. In medical research, big data analysis can assist doctors in diagnosing diseases

ISSN 2616-5805 Vol. 4, Issue 5: 54-61, DOI: 10.25236/AJMS.2023.040508

and formulating treatment plans. In the field of education, big data can help teachers understand students' learning situation and improve teaching quality. For different fields of big data, we can use different data processing and analysis methods, such as data mining, machine learning, deep learning and so on. In addition, visualization technology is also an important tool for processing and analyzing big data, which can present complex data information in an intuitive form to help people better understand and analyze the data[3].

In order to achieve the above purpose, this project example provides the following technical solution: a big data analysis method, the method includes the following steps: obtaining big data information accessed by a user during a target time period; storing the big data information in a distributed database in slices according to time; executing a data analysis task based on a predetermined set of rules for the big data information stored in the distributed database, and obtaining analysis results; configuring a data caching server to cache the analysis results; when obtaining a data query request from the front-end, calling the corresponding analysis result data as a query result in the data caching server based on the parameters of the data query request, and outputting the query result to the query result. The data cache server is configured to cache the analysis results; when the data query request from the front-end is obtained, the corresponding analysis result data is called as the query result in the data cache server according to the parameters of the data query request, and the query result is outputted to the front-end and displayed in a visualized way[4].

2.1 Algorithms are at the heart of processing and analyzing big data

In big data processing, commonly used algorithms include MapReduce, Spark, etc. MapReduce is a programming model for processing and generating big data. It divides a large data set into multiple small tasks, assigns them to multiple processors for processing, and finally combines the processing results to get the final result. Spark is a real-time computing framework, which can process large-scale data sets and provide fast iterative computing capabilities. In addition, machine learning algorithms and deep learning algorithms also play an important role in big data processing. These algorithms can find rules and patterns through automatic analysis of data, and provide support for decision-making[5-6].

2.2 Distributed database is Hbase database

As a further limitation of the technical solution of this example, the step of configuring a data caching server to cache the analysis results includes: configuring a first caching server and a second caching server; storing the analysis result data with an access number not greater than a predetermined threshold in the second caching server; and transferring the analysis result data to the first caching server when the number of accesses to the analysis result data in the second caching server is greater than a predetermined threshold to the first cache server. As a further limitation of the technical solution of this example, the step of configuring the data caching server to cache the analysis results also includes: eliminating the analysis result data with the earliest access time in the first cache server; transferring the analysis result data eliminated in the first cache server to the second cache server; and clearing the analysis result data with the earliest storage time in the second cache server[7].

2.3 Configure a data caching server for the results of the analysis

The step of configuring a data caching server to cache the analysis results also includes: eliminating the first caching server with the least number of accesses to the analysis results data in a preset time; transferring the analysis results data eliminated from the first caching server to a second caching server; and clearing the second caching server with the earliest storage time of the analysis results data. A big data analysis system comprises the following components: An acquisition unit, responsible for obtaining big data information accessed by users during a target time period. A storage unit, designed to store big data information in a distributed database in slices according to time. An execution unit, tasked with executing data analysis based on a predetermined rule on the stored big data information in the distributed database, generating an analysis result. A cache unit, which analyzes the stored big data information in the distributed database, as well as the analysis result data. This system offers a comprehensive solution for capturing, storing, processing, and analyzing big data, ensuring fast delivery of analysis results. Depending on specific needs and environments, additional components or functions may be required to enhance system performance and adaptability. The caching unit is used to configure a data caching server to cache the analysis results; and the output unit is used to obtain the data query request from the front-end, call the corresponding analysis result data as the query result in the data caching server based on the

ISSN 2616-5805 Vol. 4, Issue 5: 54-61, DOI: 10.25236/AJMS.2023.040508

parameters of the data query request, output the query result to the front-end and visualize the display[8].

2.4 Structure of the cache unit

The caching unit includes: a configuration module for configuring a first cache server and a second cache server; a storage module for storing analysis result data with an access number not greater than a predetermined threshold in the second cache server; and a transfer module for transferring the analysis result data to the first cache server when the number of accesses to the analysis result data in the second cache server is greater than a predetermined threshold. To the first cache server, a computer device includes a memory, a processor, and a computer program stored in the memory and operable on the processor, and the processor executes the computer program to realize the steps of the method. The processor executes the computer program, and the steps of the method are realized when the computer program is executed by the processor[9].

3. The manner in which the project is realized

In this project example, by obtaining the big data information accessed by users during the target time period; and storing the big data information in the distributed database by time slice; then based on the predetermined rules for the big data information stored in the distributed database to perform the data analysis task and get the analysis results; configure the data caching server to cache the results of the analysis; and in the front-end data query request, based on the parameters of the data query request, call the corresponding analysis results in the data caching server as the query results, and output the query results to the front-end and visualization display, which can reduce the burden on the processor of the big data analysis system and visualize the display. When obtaining data query requests from the front-end, based on the parameters of the data query request, the corresponding analysis results in the data cache server are called as the query results, and the query results are output to the front-end and visualization. This can reduce the burden on the big data analysis system processor, improve user access speed, avoid access lag, and ensure smooth user access.

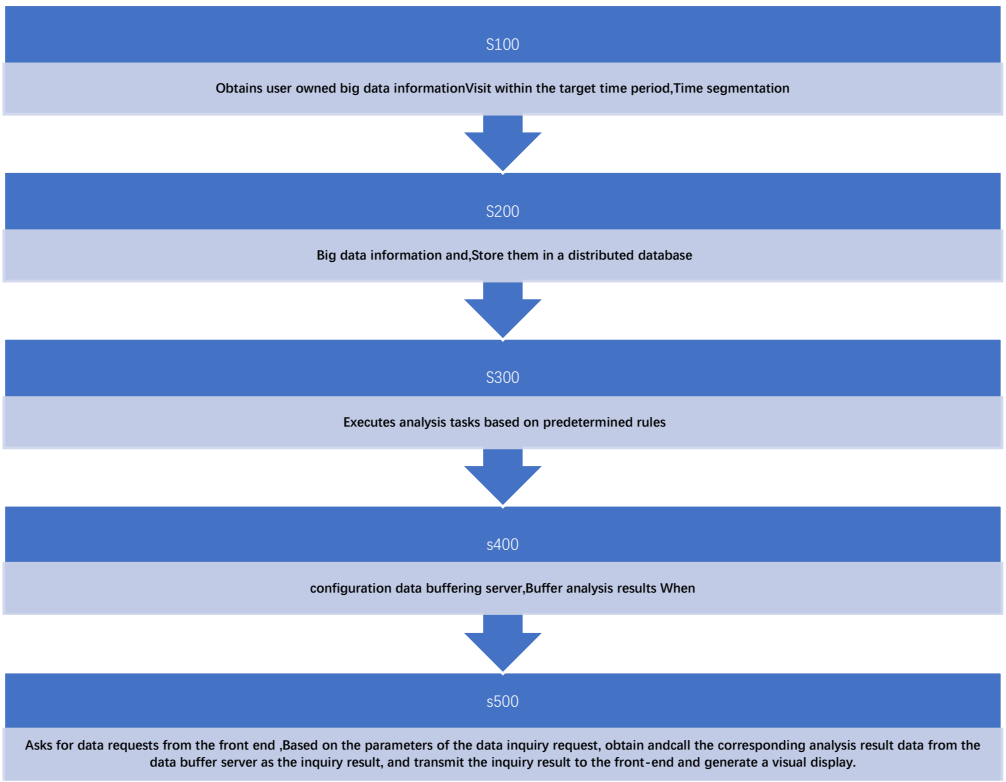

Figure 1 Architecture of Big Data Analysis Methods

As shown in Figure 1, the system architecture can include terminals, distributed databases and cache servers. Users can use the terminal to interact with the cache server through the network to receive or send messages. The terminal can be hardware or software. When the terminal is hardware, it can be

ISSN 2616-5805 Vol. 4, Issue 5: 54-61, DOI: 10.25236/AJMS.2023.040508

various electronic devices with communication functions, including but not limited to smart phones, tablet computers, e-book readers, MP3 players, MP4 players, laptop computers and desktop computers. When the terminal is software, it can be installed in the electronic equipment listed above. It can be implemented as multiple software or software modules, or as a single software or software module. No specific restrictions are made here.

It should be noted that the caching server can be hardware or software. When the server is hardware, it can be realized as a distributed server cluster composed of multiple servers or as a single server. When the server is software, can be realized as multiple software or software modules, can also be realized as a single software or software modules. There is no specific limitation here[10].

It should be understood that the number of terminals and servers in FIG. 1 is merely illustrative. There may be any number of terminals, networks and servers, depending on the needs of the realization.

3.1 The project example provides a methodology for analyzing big data

A flow of an example of a method for analyzing big data is illustrated. This example is mainly illustrated by applying the method to an electronic device having certain computing power. A method for analyzing big data, including the following steps.

Step S100: Get the big data information accessed by users in the target time period; In step S100 provided by the project example, when a user uses a terminal to search and access, the data that the user searches and accesses within the target time period is obtained through the terminal, and the time node and text information of the data are stored, and the stored time node information and text information are sent through the network. A network may be a medium for providing a communication link between a terminal and a server. The network can include various connection types.

Step S200: store the big data information in the distributed database in time slices; In step S200 provided by the project instance, the time slice can be set to one week as required, and the big data called by the server will be overwritten by the new big data one week later, so as to update the big data. In this big data analysis method, before storing the data in the distributed database, it also includes the integrity verification and legitimacy verification of big data.

Step S300: Execute the data analysis task on the big data information stored in the distributed database based on the predetermined rules, and get the analysis results.

Step S400: Configure the data cache server to cache the analysis results.

Step S500: When obtaining the data query request from the front end, call the corresponding analysis result data in the data cache server as the query result according to the parameters of the data query request, and output the query result to the front end for visual display. In step S500 provided in the project example, the visualization display uses a terminal display screen to display the output query results, so that users can obtain the query results.

In the preferred implementation mode provided by this project, the distributed database is the Hbase database, which stores big data in the way of row key and column name. In this big data analysis method, before the data is stored in the distributed database, it also includes the integrity verification and legitimacy verification of big data. The integrity verification is completed by Redis in the network system. After passing the integrity verification, the big data is sent to the server to complete the legitimacy verification. Among them, Redis is an open source log and key value database in the network system, which supports the network and can be memory based or persistent.

3.2 Big data analysis methods provided

Flow diagram of step S400 of configuring data cache server to cache analysis results, wherein step S400 of configuring data cache server to cache analysis results includes:

Step S401: Configure the first cache server and the second cache server;

Step S402: store the analysis result data whose access times are not greater than the preset threshold in the second cache server;

Step S403: When the number of accesses to the analysis result data in the second cache server is greater than the preset threshold, the analysis result data will be transferred to the first cache server.

In the above step S400 of configuring the data cache server to cache the analysis results, two cache

ISSN 2616-5805 Vol. 4, Issue 5: 54-61, DOI: 10.25236/AJMS.2023.040508

servers are configured to store the cache data with more and less accesses, respectively. The two cache servers adopt independent elimination strategies to eliminate data, which can avoid inaccurate judgment of a single cache server and eliminate some data expected to be cached. This effectively improves the accuracy of cached data.

3.3 Analyze the results of configuring the data caching server

Discloses the flow diagram further explained by step S400 of configuring data cache server to cache analysis results in the big data analysis method provided by the project example. The step S400 of configuring the data cache server to cache the analysis results also includes:

Step S4401: Eliminate the analysis result data with the earliest access time in the first cache server;

Step S4501: transfer the eliminated analysis result data from the first cache server to the second cache server;

Step S4601: Clear the analysis result data stored for the earliest time in the second cache server.

In a preferred example, data elimination is performed when the data storage of the first cache server is full. Among them, obtain a preset time, the first cache server in the access time from the current furthest data, and will be preset in the access time of the earliest data is prioritized to be eliminated, eliminated data can be transferred to the second cache server. Thus, some data with earlier access time but lower frequency of recent access will not be directly eliminated, which avoids misjudgment, and can avoid inaccurate judgment of single cache server and elimination of some data which are expected to be cached, thus effectively improving the accuracy of cached data.

3.4 Analyze the results of the configuration data slowing down

In the big data analysis method provided by the project example, the flow diagram of step S400 that further explains how to configure the data cache server to cache the analysis results. The step S400 of configuring the data cache server to cache the analysis results also includes:

Step S4401: Eliminate the analysis result data with the least number of accesses in the preset time in the first cache server;

Step S4502: transfer the eliminated analysis result data from the first cache server to the second cache server;

Step S4602: Clear the analysis result data stored for the earliest time in the second cache server.

Specifically, in a preferred example, the first cache server performs data elimination when the first cache server's data storage is full.

Among them, to obtain a preset period of time, the first cache server in the number of times the data is accessed, and the preset period of time the minimum number of accesses to the data is prioritized to be eliminated, eliminated data can be transferred to the second cache server. Thus, certain data with a high number of accesses but a low frequency of recent accesses will not be directly eliminated, which avoids misjudgment, and can avoid inaccurate judgment of a single cache server, which eliminates some data that are expected to be cached, and thus effectively improves the accuracy of the cached data, this project example provides a big data analysis method by obtaining the big data information accessed by users during the target time period; and storing the big data information in the distributed database by time slice.

3.5 The structure of the big data analytics system

Specifically, the big data analysis system 600 includes: acquisition unit 601, the acquisition unit is used to obtain the big data information accessed by the user during the target time period; storage unit 602, the storage unit is used to store the big data information in a distributed database by time slice; in this project example, the time slice can be set to a week as needed, and after a week, the big data called by the server will be covered by new big data, thus realizing the big data update. In this project example, the time slice can be set to one week as needed, and the big data called by the server will be covered by the new big data after one week, thus realizing the update of big data. In this big data analysis method, before storing the data in the distributed database, it also includes the integrity verification and legitimacy verification of the big data.

ISSN 2616-5805 Vol. 4, Issue 5: 54-61, DOI: 10.25236/AJMS.2023.040508

In the preferred implementation mode provided by this project, the distributed database is the Hbase database, which stores big data in the way of row key and column name. In this big data analysis method, before the data is stored in the distributed database, it also includes the integrity verification and legitimacy verification of big data. The integrity verification is completed by Redis in the network system. After passing the integrity verification, the big data is sent to the server to complete the legitimacy verification.

Execution unit 603, the execution unit is used to perform a data analysis task based on predetermined rules on the big data information stored in the distributed database, and obtain an analysis result.

Cache unit 604, a cache unit used to configure a data caching server to cache analysis results; And, the output unit 605 is used to obtain front-end data query requests. Based on the parameters of the data query requests in the data cache server, the corresponding analysis result data is called as the query result, and the query result is output to the front-end and visualized for display.

3.6 Big Data Analytics System Architectural Framework

A structural framework for a cache unit 604 in a big data analytics system is provided. Wherein the cache unit 604 comprises.

Configuration module 6041, the configuration module for configuring a first cache server and a second cache server; the

Storage module 6042, the storage module is used to store in the second cache server the analysis result data whose number of accesses is not greater than a preset threshold; and, the

Transfer module 6043, the transfer module is used to transfer the analysis result data to the first cache server when the number of accesses to the analysis result data in the second cache server is greater than a predetermined threshold.

Cache unit 604 is configured with two cache servers to store the cached data with more and less accesses respectively. The two cache servers adopt independent elimination strategy for data elimination, which can avoid the inaccurate judgment of single cache server and eliminate the data that are expected to be cached, thus effectively improving the accuracy of the cached data.

4. Storage media are provided

The project also provides computer equipment, the computer equipment comprising a memory, a processor, and a computer program stored in the memory and runnable on the processor, the processor executing the computer program to implement the steps of the big data analysis method.

The project also provides storage media, the storage media stores a computer program, the computer program is executed by a processor to realize the steps of the big data analysis method.

A computer program may be divided into one or more modules, one or more of which are stored in memory and executed by a processor to accomplish the project. One or more modules may be a series of computer program instruction segments capable of performing a specific function, which instruction segments are used to describe the execution process of the computer program in the terminal device. For example, the above-described computer program may be partitioned into units or modules of the berth status display system provided in each of the above-described system examples.

The above description of the terminal device is only an example and does not constitute a limitation of the terminal device, which may include more or fewer components than the above description, or a combination of certain components, or different components, for example, it may include an input/output device, a network access device, a bus, and the like.

The processor can be a central processing unit, but also other general-purpose processors, digital signal processors, special integrated circuits, off-the-shelf programmable gate arrays or other programmable logic devices, discrete gates or transistorized logic devices, discrete hardware components, etc. The general-purpose processor can be a microprocessor or the processor can be any conventional processor. General-purpose processor can be a microprocessor or the processor can be any conventional processor, etc., the above processor is the control center of the above terminal equipment, the use of a variety of interfaces and lines to connect the various parts of the user terminal.

The memory may be used to store computer programs and/or modules, and the above processor

ISSN 2616-5805 Vol. 4, Issue 5: 54-61, DOI: 10.25236/AJMS.2023.040508

realizes various functions of the above terminal device by running or executing the computer programs and/or modules stored in the memory, and calling the data stored in the memory. The memory may mainly include a storage program area and a storage data area, wherein the storage program area may store an operating system, at least one application program required for a function (e.g., an information collection template display function, a product information release function, etc.), etc.; the storage data area may store data created according to the use of the berth status display system (e.g., product information collection templates corresponding to different product categories, product information required to be released by different product providers, etc.). (e.g., product information collection templates corresponding to different product categories, product information to be released by different product providers, etc.) and so on. In addition, the memory may include high-speed random access memory, and may also include non-volatile memory, such as hard disk, memory, plug-in hard disk, smart memory card, secure digital card, flash memory card, at least one disk memory device, flash memory device, or other volatile solid-state memory device.

The module/unit integrated in the terminal equipment may be stored in a computer-readable storage medium if it is realized in the form of a software functional unit and sold or used as a separate product. Based on this understanding, the project realizes all or part of the modules/units in the above example system, and can also be accomplished by a computer program to command the relevant hardware, the above computer program can be stored in a computer-readable storage medium, the computer program when executed by the processor, can realize the functions of the above example system.

Computer program includes computer program code, which can be in the form of source code, object code, executable file or some intermediate forms. The computer-readable medium may include any entity or device capable of carrying computer program code, this project example provides a big data analysis method by obtaining the big data information accessed by users during the target time period; and storing the big data information in the distributed database by time slicerecording medium, U-disk, mobile hard disk, magnetic disk, optical disk, computer memory, read-only memory, random access memory, electric carrier signal, telecommunication signal, software distribution medium, etc.

In summary, the big data analysis method and the big data analysis system provided in this project example obtain big data information accessed by users during a target time period; store the big data information in a distributed database in slices according to time; then perform data analysis tasks based on predetermined rules on the big data information stored in the distributed database to obtain analysis results; and configure a data caching server to cache the analysis results; and when obtaining a data query request from the front-end, call the corresponding analysis result data in the data caching server as the query result based on the parameters of the data query request, and output the query result to the front-end and perform visualization. And when obtaining the front-end data query request, based on the parameters of the data query request, call the corresponding analysis results in the data cache server as the query results, and output the query results to the front-end and visual display, which can reduce the burden of the processor of the big data analysis system, improve the user access speed, avoid access latency, and ensure the smooth progress of user access.

5. Conclusion

The application of big data has permeated every aspect of our lives, bringing convenience and efficiency to our lives and work. However, the processing and analysis of big data also faces challenges and risks. Therefore, we need to adopt appropriate methods and algorithms to process and analyze big data to ensure the accuracy and integrity of the data. At the same time, we also need to pay attention to data security and privacy protection to avoid data leakage and abuse. In the future, as technology advances and data grows, processing and analyzing big data will become even more important and necessary. We should continue to explore and research new methods and algorithms to better utilize big data to bring us more value and convenience.

This project example provides a big data analysis method by obtaining big data information accessed by users during the target time period; And store big data information in a distributed database by time slice; Then, according to the predetermined rules, perform data analysis tasks on the big data information stored in the distributed database to obtain the analysis results; Configure the data caching server to cache the analysis results; When obtaining data query requests from the front-end, based on the parameters of the data query request, the corresponding analysis result data in the data cache server is called as the query result, and the query result is output to the front-end for visual display, which can reduce the number of big data query requests. Query request, based on the parameters of the data query request,

ISSN 2616-5805 Vol. 4, Issue 5: 54-61, DOI: 10.25236/AJMS.2023.040508

calls the corresponding analysis results in the data cache server as the query results, and outputs the query results to the front end and visual display, which can reduce the burden of the big data analysis system processor, improve the user access speed, avoid access latency, and ensure the smooth progress of user access.

References

[1] Zhao Qiurong, Zhang Dongxiao. Research and Development of Intelligent Construction Sites under the Internet of Things [J]. Enterprise Management, 2022, 23-23
[2] Zheng Yu, Zhao Jianping. Unified Encoding of the Internet of Things: Based on top-level, with strong inclusiveness [J]. Chinese Automatic Recognition Technology, 2014123-124
[3] Liu Ming, Liu Qing, Dong Bingrui. On the Application of Internet of Things Technology in Modern Agriculture. Southern Agricultural Machinery, 2017, 56-57
[4] Sun Aimin, Liu Na, Yin Dan, Feng Peng. Research Progress on Respiratory Rehabilitation Management System Based on IoT Big Data [J]. Chinese Journal of Health Management, 2021, 89-99
[5] Zhang Linhong. Research on the Training of IoT Application Technology Talents in Agricultural Vocational Colleges - Taking Beijing Agricultural Vocational College as an Example [J]. Journal of Beijing Agricultural Vocational College, 2022, 45-46
[6] Zhou Ji. Application Analysis of Internet of Things Technology in Smart City Construction [J]. Future Urban Design and Operations, 2022, 89-90
[7] Ye Yun. Research on the Application of Internet of Things Technology in the Context of Big Data Era [J]. Industrial Innovation Research, 2022234-235
[8] Zeng Jianqiu, Ren Mengzhao. Looking at the Role of Internet of Things Technology in Epidemic Prevention and Control from Marx's Social Science and Technology Theory. Industry and Technology Forum, 2022, 45-46
[9] Jin Shuguang. Research on Smart Agriculture Applications Based on Internet of Things Technology [J]. Seed Technology, 2022, 45-46
[10] Zhao Gang. Research on Internet of Things Technology and Its Application in the Development of Tea Industry [J]. Fujian Tea, 2022, 98-99