Application of Big Data in Agricultural Internet of Things

Jianbo Li  
Hunan Province Cooperative Innovation Center for the Construction & Development of Dongting Lake Ecological Economic Zone, Hunan University of Arts and Sciences, Changde 415000, Hunan, China

Xiaohua Li*  
Hunan University of Arts and Sciences, Changde, Hunan 415000, China  
*Corresponding author (E-mail: 2810383532@qq.com)

Yibo Peng  
Hunan University of Arts and Sciences, Changde, Hunan 415000, China

Abstract  
China is a big agricultural country, Internet of things technology, computer technology and the development of the Internet; it provides sufficient preparation for the development of agricultural informatization and modernization. In particular, digital technology and Internet technology are widely used in the construction of agricultural modernization, and to lay a good technical support for the development of agricultural Internet of things, create conditions for realizing intelligent agriculture. This paper starts with the application status of big data in China’s agricultural Internet of things, and discusses the application mechanism of big data in the agricultural Internet of things, analyze the current big data application environment, application level and existing shortcomings in rural areas, and on this basis, put forward to strengthen the application of agricultural Internet of things big data. Measures should be taken from the aspects of platform construction, enterprise participation and personnel training, promote the development of agricultural modernization. Agricultural data collection terminal based on the Internet of things, the collected data is often a large number of redundant, this brings some problems for subsequent analysis and processing. Based on the advantage of BloomFilter, an optimization method to filter redundant data is proposed, and the feasibility analysis and process description of the system model are carried out.

Key words: Big Data, Agricultural Internet of Things, Smart Agriculture

1. Introduction

The concept of the Internet of Things was first proposed at the International Conference on Mobile Computing and Networks held in the United States. At that time, the Internet was based on the Internet, using the radio frequency identification technology, wireless communication technology and other global products built in real time to share the Internet referred to as “Internet of Things” [1]. At the World Summit on the Information Society, the definition of the Internet of Things has expanded, not just the Internet of Things based on RF technology. With the development of cloud computing, the continuous improvement of big data technology and the massive data processing technology based on distributed computing platform provide ideas and methods for solving modern agricultural problems [2]. In China, the development and application of Internet of Things technology is also in a leading position in China, and has been widely used in major cities such as Beijing, Shanghai, and Zhejiang. At the 2nd China Mobile Government Symposium held by Peking University in China, the development of mobile communication and Internet of Things technology represented the formation of a new generation of information technology, which changed the social form and promoted the development of social informationization [3]. The world has attached great importance to the Internet of Things technology, and has proposed related IoT plans including the “Smart Earth” strategy proposed by the US government, the “European Internet of Things Action Plan” proposed by the European Commission, and the “Internet of Things Infrastructure” proposed by the Korea Communications Commission. Build a basic plan and so on [4]. In the same year, Premier Wen Jiabao put forward some views and requirements on the development of the Internet of Things when he visited Wuxi. China began to regard the Internet of Things as an important development strategy for China in the future [5]. The State Development and Reform Commission, jointly approved by the State Council, and the Ministry of Industry and Information Technology drafted the “Guiding Opinions on Promoting the Orderly and Healthy Development of the Internet of Things”, which was approved by the State Council. The main suggestions are to speed up the sensor network, intelligent terminals, big data
processing, and intelligent analysis, and other key technologies for research and development and innovation, and promote the integration of Internet of Things and next-generation mobile communications, cloud computing [6].

However, there are still many problems in the application of big data in the agricultural Internet of Things. On the one hand, the data platform in the agricultural Internet of Things has not yet been built. The Internet of Things, as a product of informationization in China, has been developing for about 9 years. The sensors that are available have problems such as insufficient performance, high product price and relatively poor stability, and there is no complete closed loop for sensing transmission and control. In addition, the agricultural IoT big data platform requires a large amount of network information technology as a support. There are many information barriers between regions, and it is difficult to establish a big data platform for information sharing between regions. On the other hand, the main body of big data used in the agricultural Internet of Things is the lack of professional literacy of farmers, and the lack of professional and knowledge in the application of big data.

Bo Y proposed that with the rapid development of Internet of Things technology, its application in the agricultural supply chain is becoming more and more mature, and the traditional agricultural supply chain risk is effectively controlled and evaded [7]. China’s agricultural supply chain is in a period of transition. The use of new technologies and new supply chain models can not only enhance the value of the agricultural supply chain, reduce the risks of traditional supply chains, but also bring new value to the agricultural supply chain. Therefore, the Internet of Things has important theoretical and practical significance in agriculture. The quality and safety of agricultural products has become a bottleneck restricting the sustainable development of China’s agricultural products import and export trade. In order to cope with agricultural product risks and improve service levels, Pan J put forward higher requirements for the traceability platform of the agricultural Internet of Things. The quality of agricultural products may occur in all aspects of agricultural product supply, including farming, processing, distribution, and sales. Therefore, to ensure the quality and safety of agricultural products and eliminate potential safety hazards, not only policy support, but also IoT technical support is required [8]. However, to achieve the agricultural Internet of Things, there is a need to address the efficient storage and reasoning of massive heterogeneous sensor data collected from various sensing devices. Wang explored the agricultural Internet of Things architecture based on heterogeneous sensor data and proposed a cloud computing-based agricultural Internet of Things implementation. This design is based on HBase’s two-tier storage structure. HBase is a distributed database with high scalability. It uses the distributed programming framework MapReduce model to access the database. Therefore, this design provides scalable storage, efficient data access, and simplifies other processing of sensor data [9]. Through the Internet of Things, cloud computing and big data to monitor agricultural environmental parameters, Dinesh E proposes to influence the use of different advancements like huge work information, which will provide a variety of management for agronomists, including production management, advertising, back-office management, online Business, through cloud computing network benefits, etc., will also reduce youth unemployment [10]. It has also made agriculture a necessity for people’s lives and made it a global production area, thereby further increasing gross domestic product. If this framework is implemented by ranchers, then the people and the country’s economy will grow considerably.

Big data has been seen as an important technological change, and modern information technology in the Internet of Things (IOT) and cloud computing has become a cutting-edge technology in data mining and smart applications. Therefore, the development of science and technology is Enter the era of big data. This paper first introduces the background of big data, introduces the relationship between big data and the Internet of Things, basic connotation and key technologies, and then analyzes the demand of big data in the main application areas and its impact on intelligent agriculture [11]. In response to the expansion of the overall management scope of the agricultural sector and the rapid increase in the amount of data, the agricultural system has been continuously improving the requirements of the data calculation and analysis software in the agricultural sector, and the original system has not been able to meet the needs of the agricultural sector very well. The agricultural data management and processing platform based on cloud computing platform is realized to realize the safe and reliable storage management of agricultural data, and try to use our developed algorithm to analyze and process large amounts of data, and make full use of the value brought by big data. Based on the analysis of requirements, this paper uses Hadoop cluster to build a platform for storage and management of big data, making full use of its high reliability, high fault tolerance, scalability, etc., for the distributed processing of agricultural data, introduces the research background and Significance, big data technology, Internet of Things technology and cloud computing platform domestic and international research status [12], and the reasons for the development of agricultural modernization and the use of key technologies such as Internet of Things and big data in agriculture. Then introduced two ways to build, one is built on a common computer cluster, one is built on the HPC server, and supports the difference between the computer cluster construction and the platform built on the server [13], and then tested the built platform, including database read and writes tests on the server, interactive testing with traditional database data, and synchronization on the server. Finally, it introduces how to carry out
agricultural IoT data acquisition based on wireless sensor network, wireless sensor network architecture and principle, and the design and implementation of each functional module of the platform [14].

2. Platform Design

2.1. Agricultural Internet of Things Data Characteristics and Functional Requirements

The agricultural Internet of Things is the connection of the objects related to agricultural products, so the characteristics of the agricultural Internet of Things also have a lot to do with the characteristics of agricultural data. We understand agricultural big data and its characteristics. Agricultural production data has a large number, complex structure, various forms, real-time changes and important information such as important big data characteristics, resulting in its collection, transmission, storage and management and clustering decisions. Adding great difficulties to people is mainly reflected in the following aspects: First, the amount of data generated by the agricultural production process is huge. Second, the types of agricultural production data are diverse. Third, the value of agricultural production process data is low, and it needs to be analyzed by big data technology to get value. Fourth, the processing speed is fast. With the development of agricultural modernization, more and more types of data need to be collected, and the amount of data is getting larger and larger. The general database platform used has certain limitations. With the development of cloud computing and big data technology, we handle it for us. Massive agricultural data provides ideas. At present, big data technology is developing rapidly, and some applications are quite mature, which lays a solid foundation for us to use the platform for data storage processing [15].

According to the 40th report of “China Internet Development Statistics”, as of June 2017, the number of Internet users in China exceeded 750 million. The proportion of users in rural areas has exceeded 26.7%. Correspondingly, the size of China’s online shopping market is still expanding, with a five-year average growth rate of 80% from 2012 to 2016. According to statistics, the online market size in 2016 is about 15% more than the retail sales of consumer goods. In this era, agribusinesses must strengthen the application of big data in the Internet of Things, with supply-side optimization as the goal. With the development of the times, we will continue to enhance the market competitiveness of agricultural enterprises.

2.2. Platform Database Design

The traditional database storage format and the distributed database storage format are different. The traditional database storage, the platform is composed of a large number of different types of sensor data, file and personnel data types, using a traditional database and a distributed database common composition, the server has Web server, mysql database server, and HBase database server. We will need to access frequently accessed data, such as the management of the person login using the mysql database, storing the personnel information in the mysql relational database, and storing the files directly in the HDFS distributed database, while a large number of different types of sensors The data is stored in the HBase database, stored in a distributed database with backup, high security, and the distributed database is suitable for storing a large amount of agricultural data, because it is not required for the data type.

The design of the database is designed according to the data type. The combination of the two database types makes it more efficient and stable. The relational database is more suitable for the collection of data in the actual production environment, and then designs the interface between the database and the distributed database, which can complete the data migration and interact with each other.

The data acquisition module mainly includes a wireless sensor network (WSN) module composed of a plurality of sensors and a ZigBee module, an image acquisition module, and a weather acquisition module. The wireless sensor module is mainly used to obtain real-time environmental data of farmland. The image acquisition module is mainly used to capture the real-time environmental conditions of the farmland [16]. The meteorological acquisition module mainly acquires a wide range of meteorological data. The data acquisition module mainly completes the collection of farmland microclimate data, mainly including the environmental data of farmland, such as temperature and humidity of air, light intensity, soil water content, and the image data of farmland [17]. Mainly to obtain environmental data using the WSN network composed of related agricultural sensors and ZigBee modules, and industrial cameras to acquire image data. In addition to the collection of microclimate environmental data and image data of farmland, there is also large-scale meteorological data collection. Meteorological data and crop growth are also closely related. Timely meteorological warning can effectively improve the speed of farmers’ response to disasters and reduce disaster losses. Increase crop yields. The GPRS communication technology and the 3G network card in the data transmission module respectively transmit the environmental data and the picture to the upper computer, and finally the upper computer monitoring platform stores and analyzes the received data [18], and presents the data to the computer user through a webpage in the form of a table, a statistical graph, a query interface, and the like. On the interface, the
user can view the environmental data and image images returned from the farmland and the data analysis results in real time through the webpage, so as to better provide management services for the farmland.

3. Experiments

3.1. Data Preprocessing

The actual data contains more dirty data, which is considered inconsistent. Data preprocessing is necessary. For example, higher integrity, lower redundancy, and fewer attribute associations. It refers to the preprocessing of data based on actual business needs. An integral part of data mining is data preprocessing. For processing and analyzing algorithms to achieve reliable output, you must provide data that is uncontaminated, has a low error rate, and has a low repetition rate. The main contents of data preprocessing are: 1) data cleaning 2) data integration 3) data reduction 4) data conversion. Data mining using the above method can not only improve the quality of mining, but also reduce the mining time. Data cleaning: Typically, padding with missing values, identification or deletion of outliers, smoothing of noise data, and “clean-up” of inconsistencies. Mainly to achieve the following goals: clear abnormal data, standardize non-uniform format, eliminate redundant and useless data, and correct large deviation errors. The large amount of data on the collection side of the agricultural Internet of Things has caused a general decline in data quality, affecting the output of subsequent processing and analysis algorithms, and consuming unnecessary system resources. Affects the output of subsequent processing and analysis algorithms and consumes unnecessary system resources. Properly performing targeted data preprocessing based on actual business needs and data characteristics is essential.

Bloom Filter is a relatively time-efficient data structure for retrieving whether a given element is contained in a given collection. It uses a bit array to represent a given collection, and can quickly determine if the given collection contains any given elements. The principle of the Bloom filter mainly includes the following two points, one is a bit array, and the other is to use k hash functions that are not related to each other. It uses an array of m bits to hold the information. In the initial state, the Bloom filter is an array of length m, and the elements corresponding to all indexes are set to zero. Then, the Bloom Filter uses k hash functions that are not associated with each other. The values of the k hash functions of the elements in any one set are in the range of $V = \{1,...,m\}$, respectively. Express a set of n elements $S = \{x_1, x_2, x_3... x_n\}$. Adding to any element of the Bloom Filter requires the following two steps. First, the K-order elements are separately calculated K times by k hash functions to obtain different k values in V [19]. Then set the value of the corresponding index in the array to 1. When we need to determine if any given element y is included in S, we still need to use k hash functions to evaluate the element y for K times to get the k value. Only if the value of all the bits of the k value flags is 1, then y is an element belonging to S, otherwise, y is considered to be an element in S as long as there is a position of 0. However, we need to know that if the position repeat is set to 1, it will only work for the first time, and the next few times will be invalid because it has been set to 1.

3.2. Spark Computing Framework

(1) Spark is a lightweight, fast and easy to use distributed parallel computing framework that focuses on processing without providing storage. Based on memory, it supports intermediate result iterative calculations and can process stream data for real-time effects. The Spark programming model has two main abstractions, RDD (Resilient Distributed Dataset) and two shared variables. RDD, an elastic distributed dataset, a special collection that supports multiple sources, fault tolerance, support for caching, parallel operations, and abstraction of memory-based cluster computing. There are five characteristics: partition, function, dependency, priority location, partition strategy. Dependencies describe a relationship that refers to the dependencies between RDD and parent RDD, with narrow dependencies and broad dependencies. The difference is that each partition of the parent RDD depends on at most one child RDD or multiple partitions. Narrow dependency description: If the RDD of the relationship can be shared by the same key, the entire operation can be performed on one cluster node without causing data mixing between the networks. Wide dependency description: The RDD of this relationship involves data mixing. Creating an RDD can use parallelizing an existing collection or loading an external dataset. RDD’s application logic is a series of transformations and execution operations. The conversion operation is lazy and only performs multiple conversion results when an Action is encountered. Common conversion operations include map, filter, flatMap, union, sample, etc. The execution operations are reducing, collect, count, and so on. Spark’s shared variables have broadcast variables and accumulators. Broadcast variables accessible to all nodes in memory are read-only cache variables. Accumulators can only be used to count and sum.

(2) Spark job scheduling

Related concepts: Task: The partitioned data operation flow is the smallest in terms of unit granularity. Task Set: A set of tasks that have no Shuffle dependencies and are associated with the Stage. Stage (scheduling phase):

1524
The scheduling phase corresponding to a task set. Job: The calculation of a single RDD execution Action, including at least one scheduling phase Application: In Spark, the job of the application Spark consisting of one or more jobs starts from the input RDD parsing. All RDD deferred operations only generate tokens, not immediate execution. The DAG Scheduler parses the code that runs the job submission into a DAG graph only when an operational processing mode appears in the chain. Different stage partitions are based on RDD Dependency, and different stages correspond to multiple Task Sets. The task scheduler interacts with the resource manager, and different deployment modes allow the resource manager to interact with the Spark cluster in different ways. The original DAG map is divided into complete phases and then enters the retrospective process. Constantly judge the dependencies, if it is a broad dependency, it will continue to divide to generate a stage. From the initial stage to the end stage, it will be divided into multiple stages. A phase can be submitted as a condition, and as long as all parent phases of the phase have been executed, the phase can be submitted. Once submitted, the Stage will accept the Task Set as an input parameter. This process means that the DAG Scheduler submits the Task Set to the Task Scheduler. The Scheduler Back end will apply for resources, apply via make offers, and schedule resources. Resource Offers. Task Set Manager is the way Task Scheduler manages Tasks. In addition, when the Task is executed, since the results returned by different types of Tasks are different, it is necessary to distinguish. The judgment is which kind of Task is the Shuffle Map Task or the Result Task. According to the output of the Task, the Result Task returns the output form: (1) If it is a small output, it will be placed in the Direct Task Result object. (2) If the specific size is exceeded (the default is about 10 MB), the Direct Task Result will be serialized on the Executor side, and the serialization result will be stored as a data block in the block manager. Then the BlockId returned by the Block Manager is placed in the Indirect Task Result object and returned to the Task Scheduler. The Task Scheduler then calls the Task Result Getter to take the BlockId in the Indirect Task Result and finally obtain the corresponding Direct Task Result through the Block Manager.

Spark SQL provides a data frame programming abstraction, which is a structured data processing module that performs SQL processing from the following aspects: Catalyst performs statement optimization processing; Spark SQL kernel obtains data from different data sources and executes queries. After that, the result is output as a data frame; Hive support is for Hive data processing. Therefore, Spark SQL has the following characteristics: wide data compatibility: you can get files from RDD, Parquet files, JSON files, etc., and make Hive compatible and retain Schema; high component protection; better performance optimization; multilingual. Spark SQL supports connection operations and associates tables with RMDB connections, but Spark uses RDDs for association operations. Spark SQL can only associate two RDDs at a time. This is different from RMDB, which can perform multiple table associations at the same time. However, since RDDs typically have multiple data partitions from different cluster nodes, a Shuffle operation is generated when the join operation is performed. Moreover, during the Shuffle operation, whether it is associated data or some extraneous data, Shuffle is executed, which brings a lot of network transmission and IO.

4. Discussion

4.1. Agricultural Big Data Analysis

Since the agricultural Internet of Things (IoT) was regarded as an important national development strategy in 2009, big data technology has played an important role in the agricultural Internet of Things. Through the investigation, as shown in Figure 1, the scale of China’s Internet of Things has shown a rapid growth trend, injecting vitality into the further development of China’s Internet.

Agricultural systems based on the Internet of Things generally function as monitoring and intelligent management in agricultural production environments. Typically, a monitoring network is formed by applying a large number of sensors and RFID identification nodes, and the information collected by the sensors and RFID is analyzed in time [20]. Improve the production efficiency of automation, intelligence and remote control. With the digitization, the informatization and the intelligent development of agricultural production, and the use of intelligent and sophisticated control technologies such as the Internet of Things, it has also brought about problems such as the explosive growth of data in agricultural production systems. Taking a small production environment using sensors and RFID acquisition devices as an example, the RFID reader pushes the identified tag information to the service processing layer via network transmission. Every day, by identifying and perceiving information, thousands of records can be identified. Faced with such a large amount of data, issues such as system scalability, analysis and processing time efficiency have become bottlenecks that agricultural systems using traditional data processing technologies cannot break through. The analytical needs of rapidly mining knowledge and information from large-scale agricultural data are not well met. The arrival of the era of big data has brought challenges to traditional data warehousing technology. But with Spark and some mature big data computing frameworks being adopted by enterprises, many successful cases were born in front of a large amount of data. In the context of understanding relevant business systems, combined with their own Spark and
agricultural IoT practices, a big data processing model based on agricultural Internet of Things was proposed and the data processing process was described. Spark’s big data computing framework has the speed of complex analysis and is easy to use. This article outlines the core technologies and design concepts of Spark.

Figure 1. Schematic diagram of China’s Internet of Things development in 2010-2019

In this paper, the degree of pest occurrence of rice aphids in a certain area from 2016 to 2019 was selected as the prediction target. The most common occurrence of rice aphids was stem borer and stem borer. The stem borer was mainly cold and rainy at 34-38 °C in Beibei. The amount of occurrence is small. If the spring is warm and dry, the amount of occurrence will increase. Due to the relatively normal temperature and humidity in spring, the mortality of wintering locust larvae is low, the occurrence of insect pests is early, and the degree of pests is high. The summer temperature is high and the rainfall is less. The drought is more affected by the development of larvae larvae. In China, the country with the largest rice growing country, the prevention of rice aphids has become particularly severe. With the artificial intelligence, as well as the rapid development of neural networks, the application of agricultural prediction has been continuously tried and innovated. Because the average temperature, minimum temperature, rainfall, and sunshine time are closely related to the occurrence of pests, this paper collects these four meteorological factors as the main analysis data. In order to avoid numerical problems caused by different orders of magnitude of unnecessary input variables, the input variables of the artificial neural network are normalized [21]. The purpose of normalization is to change each set of data to a number between -1 and 1. The normalization functions in Matlab are mainly mapminmax, postmminx, and trammmx. The data obtained by normalizing the collected raw data is shown in Table 1.

<table>
<thead>
<tr>
<th>Month</th>
<th>Average temperature</th>
<th>Minimum temperature</th>
<th>Sunshine time</th>
<th>Rainfall</th>
<th>Pest rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>-0.0909</td>
<td>-0.1409</td>
<td>-0.2520</td>
<td>-0.2985</td>
<td>0001</td>
</tr>
<tr>
<td>6</td>
<td>0.4824</td>
<td>0.3845</td>
<td>0.1248</td>
<td>0.3035</td>
<td>0001</td>
</tr>
<tr>
<td>7</td>
<td>0.9582</td>
<td>0.9718</td>
<td>0.9686</td>
<td>-0.7802</td>
<td>1000</td>
</tr>
<tr>
<td>8</td>
<td>0.6645</td>
<td>0.7182</td>
<td>0.5005</td>
<td>0.0420</td>
<td>1000</td>
</tr>
<tr>
<td>9</td>
<td>0.0355</td>
<td>0.0422</td>
<td>0.0000</td>
<td>-0.3665</td>
<td>0001</td>
</tr>
<tr>
<td>10</td>
<td>-0.6225</td>
<td>-0.6622</td>
<td>-0.0624</td>
<td>-0.8796</td>
<td>0001</td>
</tr>
</tbody>
</table>

4.2. Analysis of Experimental Data

The experimental environment in this section is to build a Hadoop platform on a virtual system using the CentOS7 virtual environment created by VMWare on a computer in the lab. The Spark computing platform is deployed in Spark YARN mode, and then the Join Key filtering sampling partition optimization algorithm proposed in this paper is compared with Spark Core join. See if you can reduce the amount of data in the Shuffle in the Spark Reduce phase and use data with more data skew to monitor whether Spark has been waiting for a long time or out of memory. Execution failed.
As can be seen from the data, the change in the amount of data written and read by the two connections in the two sets of experiments is more pronounced. For each set of experiments, the experimental data set is the same, and the tilt is the same. The main reason is that there are a large number of unsatisfied connection conditions and redundant data in the data. FSAF-Join performs Bloom filter filtering on connection conditions and redundant data before connecting, reducing the amount of data of the entire Shuffle at the time of connection.

<table>
<thead>
<tr>
<th>Join method</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spark default</td>
<td>32.5min</td>
<td>42min</td>
</tr>
<tr>
<td>Filter sampling and reclassify</td>
<td>10.9min</td>
<td>12.1min</td>
</tr>
</tbody>
</table>

Figure 2. Pest raw data normalization results

Figure 3. Running time comparison chart
It can be seen from Fig. 4 that the average absolute error of the Elman neural network prediction in September is 14.09%, and the prediction accuracy is not high. In order to improve the accuracy of model prediction, DS prediction theory is used to input the prediction results of BP neural network and Elman neural network as evidence of DS evidence theory. The output of BP neural network and Elman neural network cannot be directly used as DS. The basic probability assignment function of the evidence theory needs to normalize the prediction results of the neural network in order to satisfy the conditions of the basic probability assignment function.

![The mean absolute error between the predicted value and the actual value](image)

**Figure 4.** Average absolute error between predicted and actual values

5. Conclusions

With the promotion of the Internet of Things in agricultural informatization, in order to better monitor information in the production environment, the deployment of sensors and RFID nodes is also increasing. For the pursuit of real-time monitoring, the amount of data collected by nodes will become larger and larger, and will also bring a lot of redundancy, which will bring problems to our subsequent data processing and analysis. Seriously affect the performance and quality of data processing. Such huge data is a difficult problem encountered by traditional data processing methods. Hadoop, Storm, Spark, the success of data processing, has also been applied to the processing of agricultural big data, such as the preprocessing of highly redundant data. How to stay robust and efficient in big data processing and analysis is a detail that cannot be ignored. Based on the theoretical and practical experience, this paper analyzes the data flow and data characteristics of agricultural big data based on Internet of Things. According to the mechanism of Spark processing data analysis, from the perspective of reducing data redundancy, from the perspective of large table association optimization, a data middleware based on Bloom Filter is proposed, and a Spark large table association optimization method is proposed. The main research work of this paper is as follows:

1. Based on the agricultural data collection end of the Internet of Things, the collected data is often redundant, which brings certain problems to the subsequent analysis and processing. Based on the advantages of Bloom Filter filtering data, a redundant data filtering optimization method based on Bloom Filter is proposed.
2. For Spark, the big data computing framework can perform stream data processing, which can process agricultural data streams based on the Internet of Things and meet real-time requirements. However, in the face of many two tables, join operations often show problems caused by inefficiencies and data skew. An optimization method is given.
3. Apply the above optimization method to practical applications. A system model based on Spark and Agricultural Internet of Things was designed. The feasibility analysis and process description of the system model were carried out. The above three aspects are the main research contents of this paper, and also the optimization points of this paper for the application scenarios. Through theoretical analysis and experimental verification, the feasibility of the optimization point proposed in this paper is proved.
Acknowledgements

This work was supported by Key project of scientific research of Hunan university of arts and sciences [2018, 10], General project of Hunan department of education [18C0745], Research project on teaching reform in colleges and universities of Hunan province [2018, 543].

References