Research and Application of Spark Platform on Big Data Processing in Intelligent Agriculture of Jilin Province

Siwei Fu, Guifen Chen(✉), Shan Zhao, and Enze Xiao

Jilin Agricultural University, Changchun 130118, China
465136727@qq.com, guifchen@163.com

Abstract. Aiming at the demand of real-time massive data processing of Intelligent Agriculture in Jilin Province, this paper studies the big data processing of Intelligent Agriculture in Jilin Province based on Spark platform by acquiring real-time data through monitoring platform. This study first conducted the performance comparison experiment of Hadoop and Spark data processing platform, then used the Spark distributed cluster computing platform, real-time processing the big data of monitoring area. The experimental results show that the Spark platform speeds up 11.4 times faster than the Hadoop platform in the case of 100 million data sizes; and based on the Spark platform for real-time processing of big data intelligent agricultural monitoring network, not only provides memory calculations to reduce IO overhead, but also the results are faster and more accurate. The research results provide strong support for the implementation of precision agriculture technology in intelligent agriculture.

Keywords: Spark · Big data processing · MapReduce
Intelligent Agriculture in Jilin Province

1 Introduction

China is a large agricultural country, facing the ever changing agricultural production technology, Especially the introduction of remote sensing, geographic information technology, precision agriculture and expert system in the field of farmland management. This condition leads to the original variety of agricultural ecosystems turning more complex. Among them, the agricultural data gradually evolved from single quantitative transformation into qualitative transformation. The characteristics of data are massive, dynamic, spatio-temporal, etc., accompanied by structural changes. Thus, if only relying on the existing data mining technology and relational database, it has been unable to meet the current data storage and data analysis needs. It is particular important how to effectively analyze and deal with massive data, and according to the needs of the results back to the user's hands is particular important. One of the starting points in the study of big agricultural data is how to dig out valuable knowledge from complicated agricultural data. Therefore, it is urgent to use a method to classify and extract large amounts of data to find out the correlation and underlying pattern between them [1], establish the platform for analysis and preprocessing data, monitor agricultural production process, and

Published by Springer Nature Switzerland AG 2019
D. Li and C. Zhao (Eds.): CCTA 2017, IFIP AICT 546, pp. 1–12, 2019.
https://doi.org/10.1007/978-3-030-06179-1_1

combine with data mining technology to obtain real-time information, formulate appropriate coordination measures, and finally realize the precise operation of the farmland management area. Improving the quality of agricultural products under the premise of sustainable development provides a good development environment and material basis for other areas and even the whole national economic development [2].

Aiming at demand for real-time processing of massive intelligent agricultural data, this study firstly accesses the soil air temperature and humidity data from intelligent monitoring network of the National Spark Program "Integration and demonstration of corn precise operation technology based on Internet of things" demonstration area at Nong'an county in real time through the intelligent agricultural monitoring platform, proposes the Spark distributed cluster computing platform. Secondly, conducted the performance comparison experiment of Hadoop and Spark data processing platform, then used the Spark distributed cluster computing platform with extended MapReduce computational models, real-time processing the big data of monitoring area. Spark distributing cluster computing platform has efficient supported for multiple computing modes. Spark with the main characteristics of computing in memory can be applied to a wide variety of distributed platform scenarios, including batch processing, iterative algorithms, interactive queries, stream processing, and so on [3, 4]. These different calculations are supported by a unified framework. Spark allows us to integrate various processing processes simply and at a low cost, combined Spark streaming, SQL, MLlib, Graph X and other modules and data mining technology to make appropriate analysis of data processing [5], it provides the basis for the analysis of mass data in the future. The research results provide strong support for the processing of agricultural big data and the implementation of precision agriculture technology in intelligent agriculture. Thus, it contributes to the sustainable development of agriculture, and realize the modern development of intelligent agriculture.

2 Data Sources and Research Methods

2.1 Data Sources

The study area is Nong'an County of Jilin Province, located in the hinterland of Songliao plain, and it is the central part of Jilin Province, attached to the city of Changchun, located 60 km northwest of Changchun City, east longitude 124° 31′–125° 45′, north latitude 43° 55′–44° 55′, the south is adjacent to the suburbs of Changchun, the east borders on Dehui, northeast across the river and to Fuyu, north of the former Guoer Ross Mongolian Autonomous County, west of Changling, southwest border with Gongzhuling City. County 114.7 km long from north to south, east and west 97.7 km wide [6], the total area is 5400 km^2. In 2013, the county planting area of agricultural land reached 366,000 hectares, the coverage rate of fine varieties is above 98%, it is one of the important commodity grain bases in Jilin Province. Nong'an County annual average temperature of 4.7°, frost free period of 145 days, rainfall of 507.7 mm, the effective accumulated temperature of 2800°. Flat terrain, four distinct seasons, is a temperate continental climate. The soil in Nong'an county is divided into 10 soil types and 20 sub types, 50 genera and 111 species of soil. The zonal soil is black soil, chernozem [7]. This data comes from the seven real-time data monitoring stations located in Nong'an County

Kai'an Town, Hualong Town Chenjiadian Village, real-time monitoring network for intelligent agriculture, monitoring and collection of crop production environment information using wireless sensor networks, thus to realize the real-time processing and analysis of the big data of the intelligent agriculture.

2.2 Research Methods

2.2.1 Spark Cluster Computing Platform

Spark was originally born at the APM laboratory at the University of Berkeley. It is a fast, general purpose engine that can be used in large-scale data processing [8], today is one of the top open source projects under the Apache Software Foundation. Just as its name Spark, such as lightning fast cluster computing platform, the original design goal of Spark was to make data analysis faster—Not only is it fast, but it also has to be able to write programs quickly and easily. In order to make the program run faster, Spark provides memory computing, reducing the IO overhead in iterative computation. In order to make the program run easier, Spark is written in a concise, elegant Scala language [9], Scala provides an interactive programming experience. From enterprise, medical treatment, transportation to retail trade, the big data solutions offered by Spark are pushing ahead with the insights of business that have never been seen before, and thus accelerated decision-making.

The Spark project contains more than one tightly integrated component, the core of Spark is a computing engine that consists of scheduling, distributing, and monitoring applications that are composed of many computing tasks, running on multiple work machines, or a computing cluster. As shown in Fig. 1, Spark is a large and unified software stack, including Spark SQL, Spark Streaming, MLlib, GraphX, Spark Core and Independent scheduler, YARN, Mesos Modules [10]. Spark Core implements the basic functions of Spark, including task scheduling, memory management, error recovery, storage system interaction and other modules [11]; Spark SQL is a program package that

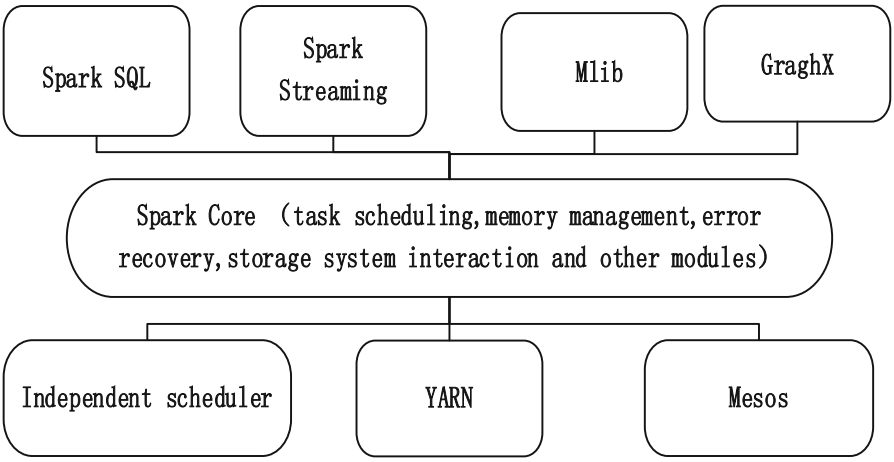

Fig. 1. Spark software stack

Spark used to manipulate structured data; Spark Streaming is a component that Spark provides for streaming computations of real-time data [12]; MLlib provides a variety of machine learning algorithms, all of which are designed to be easily scalable on a cluster [13]; GraphX is a library for operating diagrams, which can be computed in parallel [14].

In a distributed environment, the Spark cluster uses a master/slave architecture. In a Spark cluster, one node is responsible for central coordination, and each distributed work node is scheduled. The central coordination node is called the drive node, and the corresponding work node is called the actuator node. Drive nodes can communicate with a large number of actuator nodes, and they also operate as independent Java processes [15]. The drive node, together with all of the actuator nodes, is called a Spark application, as shown in Fig. 2:

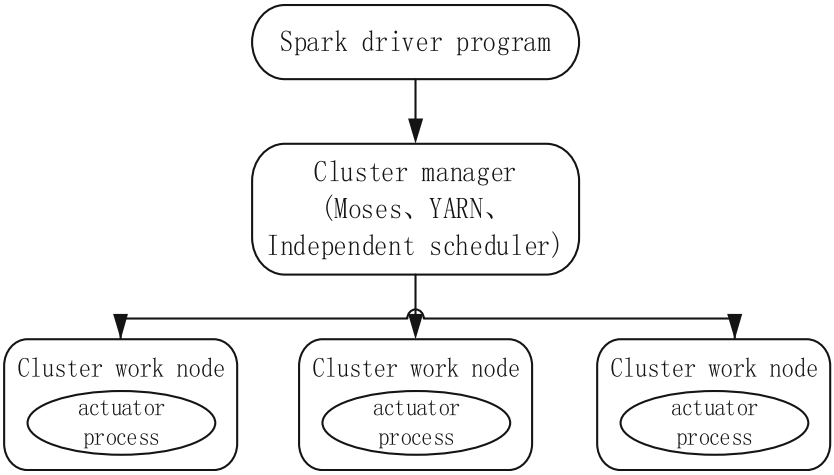

Fig. 2. Calculation framework of Spark cluster

The Spark relies on the cluster manager to start the executor node, but in some special cases it also depends on the cluster manager to start the drive node. The cluster manager is a pluggable component in Spark [16]. In this way, in addition to Spark's own standalone cluster manager, Spark can also run on other external cluster managers, such as YARN and Mesos [17].

2.2.2 The Characteristics and Advantages of Spark Platform Implementation

Although Hadoop has become the defacto standard of big data, there are still many defects in its MapReduce distributed computing model, Spark draws lessons from the advantages of Hadoop and MapReduce, and solves the problems faced by MapReduce. As shown in Fig. 3, comparing the execution flow of Hadoop and Spark can be seen, The biggest feature of Spark is the introduction of the concept of an elastic distributed data set (Resilient Distributed Dataset, RDD) [18], this allows Spark to cache data set in each node memory in cluster computing, eliminating the need to load multiple times into memory and disk storage [19], and greatly speed up the processing speed.

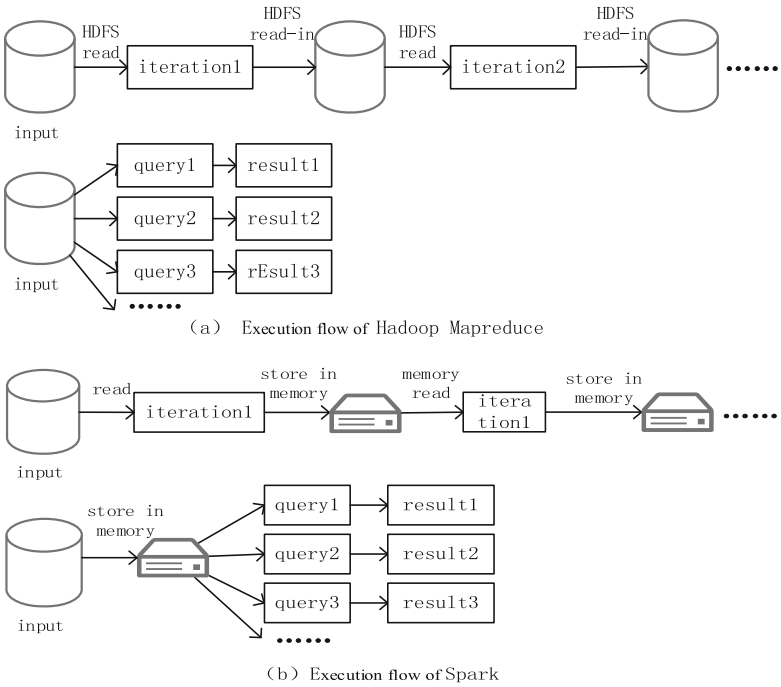

Fig. 3. Comparison of Hadoop and Spark in execution flow

About the advantages of Spark and Hadoop in executing engine, technology stack and so on, As shown in Table 1 [20–22].

Table 1. Comparison table of Hadoop and Spark performance

Type	Spark	Hadoop
Execution engine	DAG task scheduling execution mechanism	MapReduce iteration execution mechanism
Programing language	Scala, Java, Python	Java, Python, C/C++, Ruby
Technology stack	Spark streaming, SQL, MLlib, GraphX	HDFS, MapReduce, HBase, Hive
Operation mode	Independent cluster model	Distributed memory computing
Calculation model	Belong to MapReduce, but not limited to Map and Reduce, multiple data set operation types	Only Map and Reduce two operations

(*continued*)

Table 1. (*continued*)

Type	Spark	Hadoop
IO spending	Small, memory calculation	large, disk read write
Implementation of fault tolerance	RDD data storage model	Data replication
Programming code (code quantity)	Real-time interactive programming (1)	Traditional programming (3–6)
Hardware requirements	Requirements for memory and CPU	Cheap, heterogeneous
Suitable application scenarios	(1) Data mining with iterative operations (2) Real time and fast calculation (3) Machine learning operation	Delay is too high, only for offline, batch application scenarios

3 Experimental Results and Analysis

3.1 Comparison of Efficiency Test Between Spark and Hadoop

In performance, the efficiency test compares the time differences between Hadoop and Spark processing large amounts of data when performing logical regressions. This experiment from 2015–2017 intelligent monitoring network in Nong'an County 230 million soil, air temperature and humidity data in the selected 5 groups, respectively 10 thousand, 100 thousand, 1 million, 10 million, 100 million data for execution logistic regression time contrast, the efficiency test results are shown in Table 2 and Fig. 4.

Table 2. Comparison of run time between Hadoop and Spark

Computing platform	10 thousand (1 W)	100 thousand (10 W)	1 million (100 W)	10 million (1000 W)	100 million (10000 W)
Hadoop	8	17	27	255	1755
Spark	10	11	15	27	154

The number in Table 2 is the run time, units are seconds. In Fig. 4, "W" stands for "ten thousand bars".

3.2 Real Time Processing of Big Data in Intelligent Agriculture

(1) Build a stream processing framework for Spark platform

The big data analysis platform is designed to have tremendous capabilities and flexibility to meet all these requirements. The major types of processing used in big data analysis are batch processing, stream processing, and iterative processing. Therefore, we need such a platform to store such huge distributed data and perform all of these types of analysis.

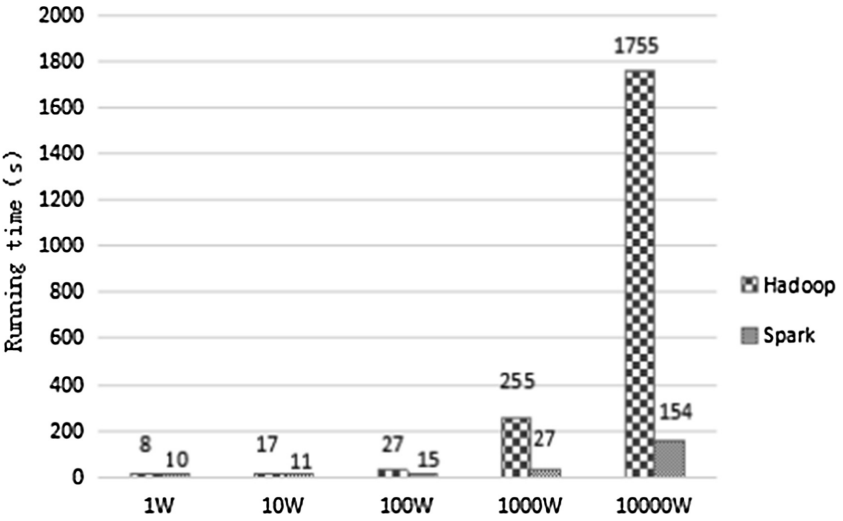

Fig. 4. Comparison of Hadoop and Spark in efficiency test

Before implementing the Spark ecosystem, you need to configure the HDFS distributed file system, build Block data blocks, Name node master nodes and Data node slave nodes; based on the HDFS distributed processing model, create HBase database to achieve high performance, real-time read and write, column storage, scalable functions; with the Thrift software framework, we build seamless and efficient service with Java, Python, C/C++, Ruby and other programming languages.

For the intelligent agriculture monitoring system in Jilin, which requires millisecond real-time response, the processing data is characterized by online, small, dynamic, relative to off-line data processing, because of the high time requirements, it is more suitable for dealing with small amounts of data and running relatively simple algorithms; for the high fault tolerance and high reliability requirement of the system, Spark uses record update to create a record RDD transform sequence, in order to facilitate the recovery of file partitions.

According to the above requirements analysis, the system uses the Spark Streaming stream computing framework with good fault tolerance and easy combination with machine learning and graph computing in the Jilin intelligent agriculture monitoring platform, real-time reception, calculation and delivery of data streams using Spark Streaming; Based on the MLlib machine learning framework, combining the improved particle swarm optimization and the limit learning machine ELM, batch processing the data for the training model; transferred processing data to the distributed file system HDFS and HBase database to store calls; the final results are fed back to the Jilin provincial wisdom agriculture monitoring platform in the form of analytical charts, as shown in Fig. 5.

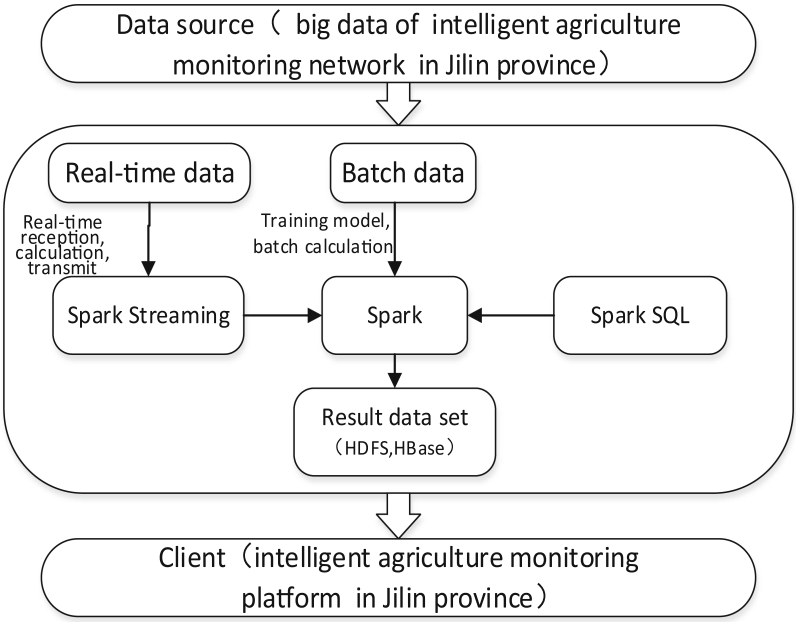

Fig. 5. Stream processing framework of intelligent agriculture monitoring platform in Jilin province

(2) Collecting intelligent monitoring network data

Based on the data comes from seven intelligent monitoring network of the National Spark Program "Integration and demonstration of corn precise operation technology based on Internet of things" demonstration area at Nong'an county, collected from May 1, 2015 to December 31, 2016 crop different dimensions (0–20 cm, 20–40 cm, 40–60 cm, 60–80 cm) soil temperature and humidity data (as shown in Table 3), and import the data into the HBase system.

Table 3. Soil moisture and temperature data of intelligent agricultural monitoring network in Jilin province

Time	Water content				Soil temperature (°C)			
	0–20 cm	20–40 cm	40–60 cm	60–80 cm	0–20 cm	20–40 cm	40–60 cm	60–80 cm
2015.5	8.75%	15.31%	19.47%	17.43%	14.51	12.63	11.18	10.11
2015.6	20.62%	20.62%	29.38%	28.85%	20.46	18.55	17.06	15.94
2015.7	12.04%	20.59%	30.73%	29.81%	24.67	23.07	21.75	20.70
2015.8	14.16%	20.91%	31.41%	30.27%	23.79	23.31	22.79	22.24
2015.9	20.69%	20.69%	30.09%	29.43%	19.02	19.34	19.47	19.41
2015.10	12.61%	19.72%	28.66%	28.13%	9.87	11.36	12.44	13.12
2015.11	5.69%	15.78%	25.63%	26.05%	2.22	4.33	5.93	7.04

(*continued*)

Table 3. (*continued*)

Time	Water content				Soil temperature (°C)			
	0–20 cm	20–40 cm	40–60 cm	60–80 cm	0–20 cm	20–40 cm	40–60 cm	60–80 cm
2015.12	3.51%	9.35%	11.63%	21.92%	−2.51	−0.30	1.38	2.48
2016.1	2.70%	7.47%	6.24%	9.91%	−8.06	−5.20	−2.70	−0.99
2016.2	4.70%	7.76%	6.15%	8.50%	−6.35	−5.31	−4.14	−3.17
2016.3	15.33%	13.83%	11.73%	11.12%	−0.75	−1.04	−0.98	−0.80
2016.4	15.85%	20.74%	25.51%	19.71%	3.96	1.39	0.09	−0.02
2016.5	15.05%	20.98%	30.96%	28.80%	12.75	10.86	7.98	8.36
2016.6	14.10%	20.92%	31.23%	29.75%	19.73	18.10	16.80	15.81
2016.7	13.05%	20.60%	31.57%	30.33%	24.08	22.73	21.58	20.68
2016.8	13.91%	22.07%	33.36%	31.36%	24.73	24.25	23.67	23.08
2016.9	18.43%	23.13%	34.19%	31.66%	18.82	19.29	19.61	19.74
2016.10	16.76%	21.51%	31.16%	29.17%	9.77	11.95	12.65	13.56
2016.11	9.18%	18.21%	27.27%	26.54%	1.83	3.79	5.51	6.78
2016.12	7.35%	9.55%	14.36%	21.40%	−2.12	0.18	1.88	3.03

(3) Processing of Soil Moisture Monitoring Data

According to Table 3 data, soil moisture data in different depths were obtained in real time, then to dynamic monitoring and multi angle data processing analysis. The processing results are shown in Fig. 6.

3.3 Results and Analysis

(1) Comparing the efficiency test of Hadoop and Spark platform

By comparing the efficiency test of Hadoop and Spark platform, the experimental results are obtained (Table 2 and Fig. 4): When the amount of monitoring data is 10 thousand, Hadoop running time is lower than Spark platform, at this point, the Hadoop is processing faster than the Spark. When the amount of data reaches 100 thousand or more, Spark takes advantage of its fast memory, computing, and distributed frameworks. In time, it is much better than Hadoop's multiple iteration algorithm. When the test content reaches 100 million, the Hadoop platform is 11.4 times more expensive than Spark. At the same time, it can be seen by the contrast chart: Under the Spark distributed framework, as the amount of data increases, the time difference between the time spent and the total time is getting smaller and smaller, reflects the stability and reliability of the Spark platform.

(2) Real-time monitoring and processing of soil moisture data

As can be seen from Fig. 6, in 2015 and 2016 in May to December year on year, the change of soil temperature at different levels is not obvious. But the soil moisture change is obvious at different levels; soil moisture content was lowest in late seedling stage (at the end of June), and it was the most in early mature (September 4th); soil temperature was lowest in the middle of seedling (June 10th), and highest in late maturing (at the end of September). At the same time, the trend of moisture change of

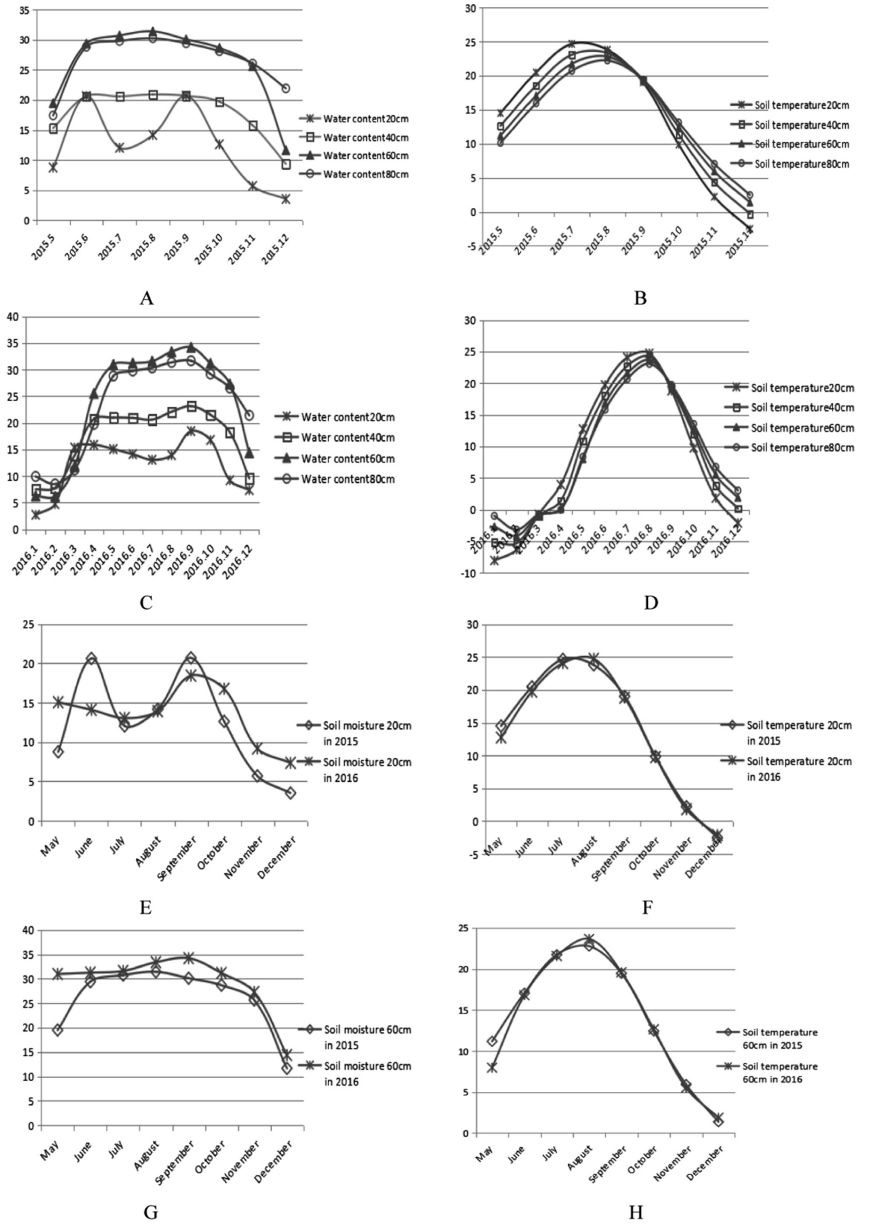

Fig. 6. Analysis of massive soil temperature and humidity data of intelligent agricultural monitoring network in Jilin province. A. Moisture trend of 0–80 cm soil in 2015; B. Temperature trend of 0–80 cm soil in 2015; C. Moisture trend of 0–80 cm soil in 2016; D. Temperature trend of 0–80 cm soil in 2016; E. Moisture trend comparison of 0–20 cm soil between 2015 and 2016; F. Temperature trend comparison of 0–20 cm soil between 2015 and 2016; G. Moisture trend comparison of 40–60 cm soil between 2015 and 2016; H. Temperature trend comparison of 40–60 cm soil between 2015 and 2016.

20 cm and 60 cm in 2015 was compared with that in 2016, affected by climate factors such as rainfall and light in different years, there were some differences, but the soil temperature did not change significantly in previous years. The analysis results show that the intelligent agriculture monitoring network based on Spark platform can deal with soil moisture data in real-time and effectively. It provides support for precision agricultural production such as timely sowing of crops and water-saving irrigation and so on.

4 Conclusion and Prospect

Through accessing the soil air temperature and humidity data from intelligent monitoring network of the National Spark Program "Integration and demonstration of corn precise operation technology based on Internet of things" demonstration area at Nong'an county in real time through the intelligent agricultural monitoring platform, research on Spark platform for big data processing of Intelligent Agriculture in Jilin province found that:

(1) Spark and Hadoop platform efficiency comparison test results show that, the Spark platform has the advantage of reducing IO overhead with its memory computing, Hadoop is more suitable for dealing with real time big data of intelligent agriculture monitoring network in Jilin province.
(2) The agricultural big data processing method of intelligent agriculture in Jilin Province based on Spark platform, using the machine learning algorithm characterized by dynamic and rapid expansion, combining the Spark streaming flow calculation framework, able to real-time analyze continuous and rapid changes in the massive data. Moreover, the calculation results are faster and more accurate. The implementation of precision agriculture and the wisdom of agricultural big data processing have a certain role in promoting.
(3) For this study, the problem of the data quantity is small, but the monitoring network real-time data processing needs large, in the future, we will further adopt the Spark platform to combine the better clustering algorithm, make the sensor transmission data to be timely and effective treatment, utilize the advantages of large data processing, combine soil moisture content with big data information, constructing perfect intelligent monitoring system of agriculture.

Acknowledgments. This work was funded by the China Spark Program. 2015GA660004. "Integration and demonstration of corn precise operation technology based on Internet of things".

References

1. Cheng, X., Jin, X., Wang, Y., Guo, J., Zhang, T., Li Guojie, J.: Large data system and analysis technology. J. Softw. (09), 1889–1908 (2014)
2. Reyes-Ortiz, J.L., Oneto, L., Anguita, D.: Big data analytics in the cloud: spark on Hadoop vs MPI/OpenMP on beowulf. Procedia Comput. Sci. **53**, 121–130 (2015)

3. Feng, Y., Huarui, W., Huaji, Z., Haihui, Z., Xiang, S.: Based on Hadoop's massive agricultural data resource management platform. Comput. Eng. **12**, 242–244 (2011)
4. Shyam, R., Bharathi Ganesh, H.B., Sachin Kumar, S., Poornachandran, P., Soman, K.P.: Apache spark a big data analytics platform for smart grid. Procedia Technol. **21**, 171–178 (2015)
5. Qi, R., Wang, Z., Huang, Y., Li S.: Based on Spark's parallel combination test case set generation method. J. Comput. Sci., 1–18 (2017)
6. Jian, L., Guifen, C., Ying, M., Hang, C.: Research and system realization of farmland environment simulation monitoring based on 3D GIS. Chin. J. Agric. Sci. Technol. **3**, 50–55 (2017)
7. Czerwinski, D.: Digital filter implementation in Hadoop data mining system. Comput. Netw., 410–420 (2015)
8. He, Q., Wang, H., Zhuang, F., Shang, T., Shi, Z.: Parallel sampling from big data with uncertainty distribution. Fuzzy Sets Syst., 117–133 (2015)
9. Yang, Z., Zheng, Q., Wang, S., Yang, J., Zhou, L.: Adaptive task scheduling strategy under heterogeneous Spark cluster. Comput. Eng. **1**, 31–35 (2016)
10. Chen, G.F., Dong, W., Jiang, J., Wang, G.W.: Variable—rata fertilization decision—making system based on visualization toolkit and spatial fuzzy clustering. Sens. Lett. **01**, 230–235 (2012)
11. Fan, Z., Zhaokang, Y., Fanping, X., Kun, Y., Zhangye, W.: Visualization of large data heat map based on Spark. J. Comput. Aided Des. Graph. **11**, 1881–1886 (2016)
12. Chen, G., Yang, Y., Guo, H., Sun, X., Chen, H., Cai, L.: Analysis and research of k-means algorithm in soil fertility based on Hadoop platform. In: Li, D., Chen, Y. (eds.) CCTA 2014. IAICT, vol. 452, pp. 304–312. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19620-6_35
13. Cai, L.: Based on large data processing technology Hadoop platform maize precision fertilization intelligent decision system research. Jilin Agricultural University, Changchun (2015)
14. Xin, W., Kan, L., Rongguo, C.: Distributed spatial data analysis framework based on Shark/Spark. Earth Inf. Sci. **04**, 401–407 (2015)
15. Xiande, Z.: Based on the Spark platform real-time flow calculation recommendation system research and implementation. Jiangsu University, Jiangsu (2016)
16. Wen, Q., Wang, J., Zhu, H., Cao, Y., Long, M.: Distributed hash learning method for approximate nearest neighbor query. J. Comput. **01**, 192–206 (2017)
17. Li, W., Chen, Y., Guo, K., Guo, S., Liu, Z.: Parallel limit learning machine based on improved particle swarm optimization. Pattern Recognit. Artif. Intell. **09**, 840–849 (2016)
18. Ziyu, L.: Big Data Technology Principle and Application. People's Posts and Telecommunications Press, Beijing (2017)
19. Yang, T., Wang, J., Yang, T., Zhang, X.: A data processing mechanism for high-efficiency large-scale graphs in Spark. Appl. Res. Comput. **12**, 3730–3734 (2016)
20. Heng, C.: A Spark-based distributed semantic data distributed reasoning framework. Comput. Sci. **S2**, 93–96 (2016)
21. Zhang, X., Chen, H., Qian, J., Dong, Y.: HSSM: a method of maximizing hierarchical data for streaming data. J. Comput. Res. Dev. **08**, 1792–1805 (2016)
22. Sun, Z., Du, K., Zheng, F., Yin, S.: Research and application of large data in wisdom agriculture. Chin. Agric. Sci. Technol. Rev. **06**, 63–71 (2013)