Received May 12, 2019, accepted July 3, 2019, date of publication July 26, 2019, date of current version August 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2928973

Fault Analysis System for Agricultural Machinery Based on Big Data

DAN LI, YI ZHENG, AND WEI ZHAO
Information and Computer Engineering College, Northeast Forestry University, Harbin 150040, China

Corresponding author: Dan Li (ld725725@126.com)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 2572018BH02, and in part by the Special Funds for Scientific Research in the Forestry Public Welfare Industry under Grant 201504307-03.

ABSTRACT With the continuous development of agricultural mechanization and information, agricultural machinery failure analysis has become an important issue. Providing effective agricultural machinery failure analysis for agricultural workers can save considerable time and cost. This paper addresses this issue by designing and producing a platform for data analysis of agricultural engineering machinery. Through the collection and analysis of agricultural engineering machinery floor data and with the help of the Internet and big data technology, physical modeling, charts, and other intuitive forms have been developed to provide information for personnel on the mechanical parts of agricultural engineering machinery, the cause of failures, and maintenance requirements. This novel agricultural engineering machinery management analysis platform currently has access to 100 million data points, with system access for capacity expansion, and provides fault detection and maintenance examples. It has already been applied in multiple farms across many of China's provinces.

INDEX TERMS Agricultural machinery, big data, sensor, data analysis.

I. INTRODUCTION

With the continuous development of modern information technology such as the Internet of Things, cloud computing, big data, mobile Internet, etc., the transformation and upgrading of the entire agricultural industry chain has become a new form of industrial revolution [1].

The United States is the first country to install systems such as satellite navigation on agricultural engineering machinery. This has opened up a new era in the use of high-tech, high-performance, and intelligent equipment in agricultural machinery. As a pioneer, the United States has nurtured many world-renowned agricultural companies, such as Monsanto Agriculture and CNH [2].

In some large farms in Europe, data exchange has been established and started using wireless communication lines between farm office computers and agricultural machinery in operation [3]. An information management system is established through these data exchanges. Through the calculation, analysis and processing of the working condition data by the system, it is possible to realize the statistics of the working area, fuel consumption rate, output, and operation. The system also displays intelligent functions such as a friendly human-machine interface. Italy's iFarming has implemented a complete range of products and services in the field of precision farming, such as the esiFARM platform. The esiFARM platform is an IT information platform based on cloud computing technology. The platform consists of four parts: data warehouse, sample transmission system, data engine system and user panel. The sampling transmission system obtains the temperature and humidity of the current planted land, the concentration of atmospheric components, the soil conductivity and even the current fruit volume of the crop through the sensors of the drone, satellite and ground [4]. The data is then passed back to the data warehouse via data networks such as LPWAN, GPRS/GSM, and Wi-Fi [5]. The data engine can establish proprietary models such as atmospheric evolution models and plant growth curves for sampling plots through a large number of parallel calculations. Finally, relevant recommendations on production are given. The irrigation system automatically changes the external input material of the sampling plot to achieve automated agricultural production.

The ''Smart Assistant'' system for agricultural machinery developed by Yanmar has been installed and promoted on agricultural machinery in Japan. The system is equipped with GPS antennas and other associated communication terminals

The associate editor coordinating the review of this manuscript and approving it for publication was Jean-Pierre Chanet.

on agricultural machinery. The system can realize visual surveillance anti-theft, operation status management, maintenance guidance service, automatic notification of sudden abnormalities and rapid response. The ''Smart Assistant'' not only supports the automation of agricultural machinery, but also works with agricultural cloud applications provided by third-party companies such as the ''Facefarm Production Resume'' program. This can increase production efficiency [6]. At present, ''Smart Assistants'' have been promoted and applied in agricultural production sites throughout Japan.

With the continuous improvement of China's informatization level, relevant management departments of agricultural machinery, colleges and universities and related commercial companies began to cooperate. By making full use of technologies such as ''internet plus'' and big data, the intelligent management of agricultural machinery has been basically realized [7]. Ningbo Agricultural Machinery Station cooperated with Ningbo China Mobile Branch to build a smart agricultural machinery information service platform. The platform organically integrates many advanced technologies such as wireless communication, agricultural machinery positioning, geographic information, and computer control. The platform can realize the functions of agricultural machinery positioning, agricultural machinery dispatching, agricultural machine working area calculation and statistics. Zhengcheng Technology Company Ltd. and Jiangsu Beidou Satellite Application Industry Research Institute jointly developed the ''Beidou Agricultural Machinery Operation Fine Management Platform'' [8]. The platform can provide intelligent and refined management services such as positioning monitoring, command and dispatch, area statistics and information management for agricultural machinery operations.

The ''Revo Apos Smart Agriculture 2025'' project jointly organized by schools and enterprises is an all-round combination of information technology and agricultural production. The project is based on the ''iFarming'' (the first smart farming solution in China) of the Revo Apos Group. Relying on the internet, internet of things and big data and other related technologies to achieve the integration and interconnection of agricultural information. Providing scientific management and business decisions for agricultural production processes by measuring basic farmland data, collecting location information, and matching intelligent agricultural machinery. This greatly improves the efficiency of the entire system. While improving productivity and improving agronomic performance, it maximizes control of production costs, improves overall production efficiency, and accelerates technological upgrading of the agricultural industry. But the project is still only in its infancy [9].

However, as most of China's existing research has focused on the synthesis of information, the statistics and records of agricultural machinery use, it is not possible to conduct data analysis and provide reasonable advice based on existing data [10]. As far as China's current situation is concerned, there is an urgent need for a solution to smart agriculture for agricultural production. The focus of this research is therefore to visualize the life cycle of components through the analysis of agricultural machinery conditions and related physical property curves, visually displaying the health of agricultural machinery and providing fault analysis and maintenance suggestions. This research not only provides comprehensive, accurate and scientific information services for China's agricultural production, but also breaks the monopoly of foreign industry giants and provides a strong information technology guarantee for large-scale mechanized production of agriculture.

The structure of this paper is as follows: the next section describes data acquisition and real-time processing; the third section provides key algorithms; the fourth section describes system parameter settings and system architecture design, as well as data storage and analysis; the fifth section conducts an experimental evaluation and displays the results; and the final section provides a summary.

II. OVERALL FRAMEWORK

This research focuses on agricultural machinery such as wheel-type tractors, cultivators, harvesters, and dryers in specific areas where production activities are carried out. First, collecting data on agricultural machinery conditions. Then, through the construction of a big data platform, the internal relationship of various indicators of agricultural machinery conditions can be analyzed. Finally, timely warning or prompting for parts loss and maintenance needs. Thereby reducing maintenance time and maintenance costs during farm hours. The main contents of this research are as follows:

- Collection of raw data
- Transmission of raw data
- Real-time monitoring and alarming of real-time monitoring systems
- Model establishment of different mechanical components
- Simple visual system interaction

III. DATA ACQUISITION AND PROCESSING

The data collected for the study is the so-called floor data of the agricultural machinery when performing agricultural operations. This floor data is the original information of the mechanical working condition with time information, which is also called the floor metadata. The data transmission process of this topic:

- The sensor collects analog signals such as temperature and speed
- The signal is converted into a digital signal by A/D conversion
- Parallel or serial transmission to the in-vehicle mobile data center via a direct connection
- The vehicle mobile digital center carries out the necessary unpacking and solving
- Send data to the IEEE 802.1 sender and mobile telecommunications network transmitter

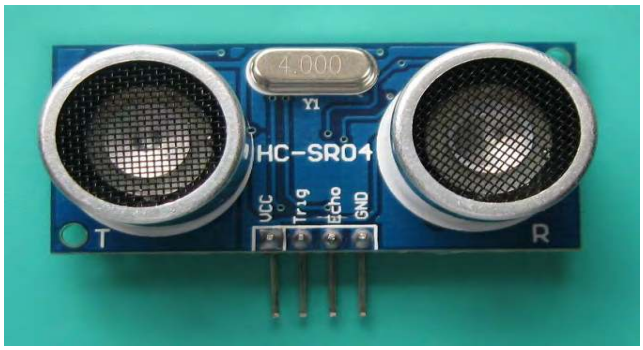

FIGURE 1. Ultrasonic sensor physical map.

• Transfer data to the remote data center of the management analytics platform

A. DATA ACQUISITION

1) RANGE DATA COLLECTION

An ultrasonic distance sensor using a micro-processing chip, as shown in Figure 1, is the distance information acquisition sensor attached to agricultural engineering machinery. An ultrasonic distance measuring sensor is a sensor that uses ultrasonic waves as a signal source. Ultrasound has a certain penetration ability. At the same time, ultrasound relies less on the measurement environment. That is to say, the ultrasonic distance sensor is more resistant to dirt and environmental resistance. Ultrasonic distance sensors have no special advantages and disadvantages, and are relatively moderate in terms of cost, accuracy, and range. The sensor is placed on the upper part of the support wheel of the equipment and is used to measure the relative distance between the main beam of the equipment and the soil surface [11]. This value can be used to obtain the working depth of the mounting equipment. By comparing this value with the height of the standing point of the lower link in the suspension system, the current horizontal state of the mounting device can be obtained. The force of the transverse section can then be decomposed so that the lifting force of the suspension device can be obtained by the laser distance sensor, and the force currently hung in each part of the longitudinal direction of the device can also be obtained.

The laser distance sensor is used as a distance collecting sensor for the relative distance between the agricultural engineering machine and the attached mounting device. It measures the relative distance between the suspension tie rod node of the hanging equipment attached to the agricultural machinery (such as the rotary tiller) and the axis of the lift arm [12]. The numerical value and the length of the suspension system, the length of the lifting lever, and the length of the top link are mathematical trigonometric functions, which can be used to obtain the actual lifting force of the current suspension system.

2) COLLECTION OF POSITION COORDINATE DATA

The displacement data collected by the position coordinate sensor can record the travel distance of the agricultural machinery during the whole operation. The actual working time and working area of the mounting device are obtained by the time series data of the distance sensor, and the auxiliary analysis model calculation. This paper describes the two types of location data acquisition methods for satellite positioning and network positioning.

Satellite positioning is a position calculation method in GPS that measures the distance between the measured object and four satellites at different positions to obtain the position data of the measured object. Formula 3-1 (electromagnetic wave transmission distance equation is used to calculate the distance between the measured object and the first satellite, where c is the speed of light constant, ΔD_i is the distance between the measured object and the i-th satellites, T_{i_0} express the time when the satellite sends a signal, and T_{i_1} indicates that the measured object has received the moment of the i-th satellite signal. The data solution of the other three satellites is the same as that of the first satellite [13].

$$C = \frac{\Delta D_i}{Ti_1 - Ti_0} \tag{3-1}$$

According to the global unique identifier of the satellite, the position of the satellite at the moment when the signal is sent can be queried in the ephemeris database of the global positioning system. Then X, Y, Z are solved according to Equation 3-2 ternary equations. X, Y, Z represent the relative coordinates of the measured object in a three-dimensional rectangular coordinate system, X_i, Y_i, Z_i represent the coordinates of the moment when the i-th satellite emits a signal in a three-dimensional rectangular coordinate system, and Di indicates the distance of the measured object and the first i-th satellites.

$$\begin{cases} (X_1 - X) + (Y_1 - Y) + (Z_1 - Z) = D_1^2 \\ (X_2 - X) + (Y_2 - Y) + (Z_2 - Z) = D_2^2 \\ (X_3 - X) + (Y_3 - Y) + (Z_3 - Z) = D_3^2 \\ (X_4 - X) + (Y_4 - Y) + (Z_4 - Z) = D_4^2 \end{cases} \tag{3-2}$$

According to the principle of network positioning and the principle of satellite positioning, if multiple satellites in satellite positioning are replaced by network base stations, the same principle can be used to obtain the position coordinate data of the measured object [14]. Table 1 provides a brief description of the respective characteristics of satellite positioning and network positioning. According to the characteristics and the actual working environment of agricultural engineering machinery, in order to meet our measurement needs, this study uses two methods of complementary positioning and data fusion.

In this design, there are two different models of position coordinate sensors. Figure 2 (physical map of the position coordinate sensor) shows the connection diagram of the PM01 sensor and the host computer. The PM01 sensor is connected to the Raspberry Pi 3B+ mainboard through a USB serial port to realize the power supply and data reading function of the PM01 module.

TABLE 1. Positioning method comparison table.

signal source	Communication satellite	Network base station
measurement method	Dynamic measurement	Static measurement
measurement accuracy	Higher	general
Measuring range	Larger	Smaller
Technical maturity	mature	More mature
Confidence	Higher	general
Environmental dependency	Higher	Lower
Power consumption	Higher	Lower
Integration difficulty	Lower	Lower

FIGURE 2. Position coordinate sensor physical map.

3) COLLECTION OF MOTION AND POSTURE DATA

The attitude data of an object is the posture of the object relative to the surface in space. Using the accelerometers and gyroscopes in the X-axis, Y-axis, and Z-axis directions, a total of six data scalar values can accurately reflect the attitude information of the object [15].

In this study, the motion attitude sensor uses the MPU-6050 microprocessor. The processor is the world's first motion-tracking microprocessor that integrates a 3-axis accelerometer and a 3-axis gyroscope. It is a DMP (Digital Motion Processor) that is directly connected via hardware. The MPU-6050 can use both IIC and SPI interface data buses. The maximum range of the gyroscope is $\pm 2000°$/sec (dps) and the fourth gear is adjustable. The accelerometer has a maximum range of ± 16g and four gears adjustable, and the on-board 1MB space queue buffer. Power consumption can be reduced by reducing data output without sacrificing accuracy. The MPU-6050 chip is packaged in a 4X4mm electronic patch with a 0.4mm solder joint exposed around it. It has a total of 24 input and output pins, 8 of which have no internal link.

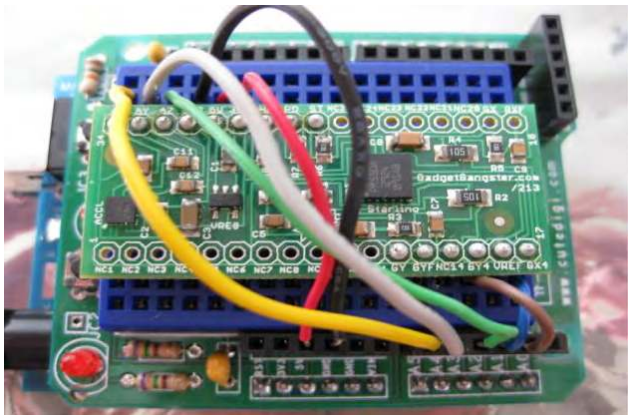

FIGURE 3. Physical connection diagram.

Because the data collected by the motion attitude sensor is motion data, it requires a good fixation between the sensor and the measured object. In this experiment, the motion attitude sensor is fixed between the expansion board and the upper motherboard. On the one hand, it is fixed by the DuPont wire pins on both sides, while on the other hand it is assisted by the active heat dissipation system of the host computer motherboard. Connection diagram shown in Figure 3.

The motion attitude sensor records all posture data of the agricultural machinery and the auxiliary mounting equipment during the running time from start-up to shutdown. The attitude data can be used to calculate the stress situation of the equipment, which is of great value to the power output shaft spindle, suspension system lift arm bearing, oil seal health and fault analysis.

4) COLLECTION OF FORCE DATA

Force type data refers to the data directly used in Newton force as the unit or the converted data in agricultural engineering machinery and its associated mounting equipment, such as oil pressure, brake system pressure, etc. At present, the acquisition of force data has a very mature solution. For example, the oil pressure can be directly added to the pressure line by adding a pressure probe to the oil pipeline, and then the current oil pressure can be obtained by the potential difference.

In this study, the force sensor uses a PCM300 pressure sensor. The probe of the hydraulic sensor is connected to the detection branch of the hydraulic pump output oil pipe through the high-pressure oil pipe, so the pressure data of the hydraulic pipe can be collected without affecting the original function of the hydraulic system. The physical picture of the sensor is shown in Figure 4.

The force type data sensor collects the hydraulic oil pipe pressure of the suspension system, the auxiliary wheel lubricating oil pressure, and the auxiliary oil pressure of the equipment. After the attitude sensor is compared with the pressure of the hydraulic oil pipes of the motion attitude, the failure analysis model of the data supply pump is used.

FIGURE 4. Hydraulic sensor physical map.

B. DATA REAL-TIME PROCESSING

The real-time processing of data is divided into two parts: the first part is data conversion, and the second part is data dynamic adjustment.

Data conversion refers to the necessary process of unpacking, solving, noise reduction, filtering, and data fusion of the data packet that the vehicle mobile data center transmits back to the lower computer. The raw data is then converted into a data format that can be used directly by the data flow of the management analytics platform. The environment in which the sensor collects data in this research is in actual agricultural production operations, and is inevitably interfered with by environmental variables, resulting in data noise. Therefore, the in-vehicle mobile data center needs to perform noise reduction on the data before sending the data to the remote data center of the management analysis platform. The data filtering process is performed by the necessary filtering algorithm (such as Kalman filtering) to ensure that the subsequent analysis process does not generate large data deviation. For specific data, there is always some deviation in the data collected by a single sensor, and data fusion of various sensors is required to achieve data reduction.

Dynamic adjustment of data refers to the in-vehicle mobile data center after data conversion to obtain directly available data. This data is copied to the asynchronous feedback data stream logic, which performs calibration according to the set threshold and the sampling parameters corresponding to different sensors. When the control parameters of the sensor are re-optimized, the re-optimized adjustment data is sent to the upper computer. The host computer writes to the sensor's control register through a direct connection, so that the next sampling of the sensor uses the latest optimization parameters. For example, attitude sensors (accelerometers and gyroscopes) have a range and allowable error limits for each accelerometer and gyroscope component. The gyroscope range has two ranges of 250° and 1000°. If the sensor sampling tolerance is ±2% of the maximum range, then when the measured data is around 200°, the test error of the 250° range is ±5 dps and the test error of the 1000° range is ±20 dps. In this comparison, the sampling parameters of the sensor sampling must be dynamically adjusted according to the actual measured data, so that the difference between the collected experimental data and the actual situation is minimized.

IV. KEY ALGORITHMS

A. DATA FILTERING

Data filtering refers to the process of filtering and rejecting interference signals with specific parameter scalar features from the original signal [16]. Because low-level hardware filtering has been considered in the design, we only need to consider software filtering under complex conditions, that is, using software algorithms for noise reduction. KALMAN FILTER, also known as Kalman filter algorithm, has optimized autoregressive data linear processing. Equation 4-1 (Kalman expression) is the most primitive expression of KALMAN FILTER. The specific mathematical derivation process and subsequent expansions and transformations are not repeated here [17].

$$X(k) = A \cdot X(k-1) + B \cdot U(k) + W(k) \tag{4-1}$$

Multi-stage filtering, also known as composite filtering, is a multi-filtering process in which the pointer performs parallel or serial processing on the same signal. In the limit-damping composite filtering method, the maximum data fluctuation difference and the maximum fluctuation counter are set. If the deviation of the current data value is within the fluctuation range, the current data is valid, the current data is recorded, and the fluctuation counter is cleared. Otherwise, the current data is discarded, the current data is replaced with the last valid data, and the fluctuation counter is incremented once. When the fluctuation counter value is greater than the maximum fluctuation counter value, the current valid value is replaced with the current data value. The limiting-debounce composite filtering method can not only remove the pulse interference caused by external factors, but also effectively avoids the data jitter caused by the frequent operation of the controller near the critical value.

B. DATA FUSION

Data fusion, also known as data integration, refers to the acquisition of new values by multiple algorithms with different sources, different information densities, and different information characteristics to obtain additional information or higher data accuracy [18]. Equation 4-2 (a complementary fusion algorithm) is one of the information fusion algorithms used in this study, where θ represents the fusion target quantity, α and ω represent the two data quantities to be fused, and k represents the fusion sequence, t representing time.

$$\theta_{\mathrm{k}} = \alpha\theta_{\mathrm{k}-1} + (1+\alpha)\,\alpha_{\mathrm{k}} + \alpha \mathrm{w}_{\mathrm{k}} \Delta \mathrm{t} \tag{4-2}$$

According to the equation (a complementary fusion algorithm), we can fuse the attitude scalar data collected by the motion attitude sensor. For example, according to the accelerometers of the three axial directions of the X, Y, and Z of the measured object and the scalar data of the angular velocity, the motion pose vector data of the measured object in the three-dimensional rectangular coordinate system can be obtained [19].

C. ANALYSIS ALGORITHM

1) TIME SERIES

Time series algorithms are targeted for data analysis with strong temporal relationships. the time column of a single key in a single sample of data is unique and continuous. there is no obvious jump in the numerical change in a continuous time range, and in each sample data besides the time column can form new keys by addition or combination [20]. when the management analysis platform performs data analysis, if it is found that data is faulty and data is missing in the data transmission, the time series algorithm can be used to write back the predicted data to eliminate the fault and supplement the data. equation 4-3 (autoregressive equation) is the starting point for time series analysis, where y_t is the current value, μ is the constant term, p is the order, γ_i is the autocorrelation coefficient, and ξ_t is the error.

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \xi_t \tag{4-3}$$

When data fluctuations are encountered, the sample data to be analyzed cannot be directly analyzed by equation 3-3 (self-regression equation) [21]. it is necessary to first use the data difference to transform the sample data to be analyzed into a smooth time series with no fluctuation or less fluctuation. the first-order difference refers to the subtraction of two sequence values separated by a time series from the original time series value. the difference equation is shown in equation 4-4 (difference equation) [22], where y represents the differential variable and y_t represents the value of the t-th time series.

$$y_t = \Phi y_{t-1} + \omega_t \tag{4-4}$$

After the samples are differentiated, a moving average of the samples to be analyzed is also required. equation 4-5 (moving average equation) is brought into the differentially processed sample data, where y_t represents the weighted average of the last two time series and ξ_t represents a white noise sequence.

$$y_t = \mu + \xi_t + \theta \xi_{t-1} \tag{4-5}$$

After the differential processing and moving average processing of the sample data to be analyzed, we have ''smoothed'' the fluctuating data into the sample data. however, in the case of differential and moving averages, we only select sample data for adjacent timings and have implemented ''local'' processing. therefore, the sample data after the above processing is finally brought into formula 4-6 (exponential smoothing equations), and exponential smoothing is used to make the fluctuating sample data to be analyzed directly available. s_i represents the smoothed value of the current time series, also represents the weighted average of s_{i-1} and the previous time series value, and t_i represents the smoothing trend term.

$$\begin{cases} S_i = \alpha x_i + (1-\alpha)(S_{i-1} + t_{i-1}) \\ t_i = \beta\,(S_i - S_{i-1}) + t_{i-1}(1-\beta) \end{cases} \tag{4-6}$$

After the four steps of auto-regression, differential processing, moving average processing and smoothing index processing with strong time relationship, regardless of whether the original data has data faults or data fluctuations, the current data to be analyzed will be a smooth curve on the waveform that achieves the output form of the time series algorithm.

2) ANOMALY DETECTION

In this study, all working condition raw data and structured working condition data have clear time stamps and have clear values. This provides an important basis for the choice of the type of algorithm model. The time series anomaly detection model is a model for analyzing time series data [23]. It can find discrete data points that deviate from the sample body based on set criteria or normal signals, including unknown expected peaks, valleys, trend changes, level changes, and so on. Additive types such as Additive Outlier, Temporal Change, Level Shift or Seasonal Level Shift [24]. In order to analyze problems similar to the above by using the raw data of the working conditions, the CART (Category Regression Tree) analysis model was adopted in this study.

Usually, there are two ways to detect anomalies. The first is that the data values for a certain period of time are manually marked as abnormal or normal. The second is to predict the normal value of a time series of data. The difference between the normal value and the actual value is compared, and the abnormality is determined according to the degree of deviation of the predicted value. The classification regression tree is derived from the second scheme, which can visually calibrate the confidence interval or the exact prediction error value. Then, through simple comparison, it is possible to obtain whether the measured data is abnormal [25]. The following outlines the application of this model in this study. Let X and Y be the input parameters and output parameters of the model, respectively, and Y is continuously changed in the time dimension. Recursive arithmetic logic is employed in the sample data set analyzed. Two sub-areas are defined in each area. A binary decision tree is constructed by determining the output values for each region. As in Equation 4-7 (2-fork decision tree).

$$\mathrm{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \tag{4-7}$$

After completing construction of the binary decision tree, the optimized variable j and the segmentation points are selected such that when the variable j is traversed, the segmentation points is scanned for the specified segmentation variable j. Select the (j, s) pair for the minimum value of Equation 4-8 (the optimal segmentation equation), where R is the divided input space and c is the fixed output value corresponding to space R.

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i \in R_i(j,s)}(x_i - c_1)^2 + \min_{c_2}\sum_{y_i \in R_i(j,s)}(y_1 - c_1)^2\right] \tag{4-8}$$

The (j, s) pair obtained by the optimal segmentation equation is used to divide the region, and the corresponding output value is determined by referring to Equation 4-9 (dividing the reference equation group).

$$R_1(j,s) = \left\{x \,\middle|\, x^j \le s\right\}$$
$$R_2(j,s) = \left\{x \,\middle|\, x^j > s\right\}$$
$$\widehat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m(j,s)} y_i$$
$$x \in R_m, \quad m = 1, 2 \tag{4-9}$$

After the division is completed, the optimization of the two sub-areas is continued, and the input space can be divided into M areas. A decision tree is then generated according to Equation 4-10.

$$f(x) = \sum_{m=1}^{M} \hat{c}_m I(x \epsilon R_m) \tag{4-10}$$

Once the decision tree is generated, the input space partition can be considered to have been determined. Then, using Equation 4-11 (square error) to represent the prediction of the analyzed data by the regression tree, the optimal output value on each sequence unit is solved by the criterion of the smallest square error.

$$\sum_{xi \in R_m} (y_i - f(x_i))^2 \tag{4-11}$$

After the above processing is completed, a confidence interval with an approximate reliability of 0.95 can be obtained. Then, the detected time series data is compared with the confidence interval. If the time series data to be tested is not within the confidence interval, the current time series data to be tested is abnormal data [26].

V. SYSTEM DESIGN

A. SYSTEM PARAMETER SETTING

The specific performance of the management analysis platform, such as processing speed, data storage capacity, network delay and sampling frequency, can be intuitively and quantitatively interpreted through the platform's performance index parameters. The four parameters are set as: response time, storage space, sampling frequency, and minimum unit.

1) RESPONSE TIME

Response time, also known as response delay, refers to the time interval between the start and end of a request (which can be a network request or a service request) or a call (which can be a function call or a system call). The response time is generally considered to consist of three parts: actual processing time, data transmission time, and performance time.

It is necessary to consider transmission of the data between the systems and the subsequent expansion of the management analysis platform to access more services to the device. Therefore, according to each management analysis platform, 10 components are expanded, and each remote data center is connected to 100 in-vehicle mobile data centers, and each in-vehicle mobile data center simultaneously accesses 100 data collection points. Since the real-time fault alarm is required to reach the second-level monitoring, the single response time is less than 1 millisecond.

2) STORAGE SPACE

In the management analysis platform of this design, the storage requirements of the raw data of the working conditions are the largest. On the one hand, the raw data of the working conditions is time series data; as long as the agricultural engineering machinery engine is running and the equipment is running, the continuous data is generated and needs to be stored and recorded. On the other hand, the raw data of the working condition is not deleted so that the storage space requirement of the original data of the storage condition will only increase and will not decrease. The storage space requirements are estimated below with a specific example.

In this design, one of the raw points in the agricultural engineering machinery working conditions is to collect the movement and posture information of the agricultural engineering machinery at work through the motion tracking microprocessor. The data is collected from the sensor, processed by the lower computer, and sent to the in-vehicle mobile data center in text format showing the working condition as per the following sample:

{"componentId": "ab61f0578d39532ab5d01a3adc0910c7",
"pin": "D_IIC",
"model": "master",
"accel_range": 16,
"gyro_range": 1000,
"accel_test": 0,
"gyro_test": 1,
"accel_filter":0,
"gyro_filter":0,
"accel_sample_rate":1,
"gyro_sample_rate":8,
"level": 64,
"timestamp": 1537410741090,
"accel_xout_h": 255,
"accel_xout_l": 255,
"accel_yout_h": 255,
"accel_yout_l": 255,
"accel_zout_h": 255,
"accel_zout_l": 255,
"gyro_xout_h": 255,
"gyro_xout_l": 255,
"gyro_yout_h": 255,
"gyro_yout_l": 255,
"gyro_zout_h": 255,
"gyro_zout_l": 255,
"temp_out_h": 255,
"temp_out_l": 255
}

This sample data is in the GPT (Globally Uniquely Identified Disk Partition Table) format disk, with NTFS (Microsoft's New File System) partition, and the cluster number is 4096, occupying 673 Bytes (bytes).

The motion attitude data of the sample must be recorded at least three times per second. It is assumed that the data collected by all the data collection points of each agricultural engineering machine is the same as the motion attitude data, and 100 vehicle movements are connected according to each remote data center. In the data center, each vehicle mobile data center is connected to 100 data collection points at the same time. Each agricultural engineering machine has an average of 10 hours of work per day (due to different types of operations, some field operations can only be carried out during the day, such as weeding operations; some can operate 24 hours a day, such as combined harvesting operations). It is 9 months of farming cycle per year. According to Equation 5-1 (storage space conversion), the management data of a single component management analysis platform raw data volume size is 2tb/year, which is the basic requirement in the ideal case of 100% disk utilization. Taking into account the actual disk utilization, data backup, analysis intermediate results, data structured storage, attached storage modules, and other factors, then when the management analysis platform equipment is fully loaded, the annual storage space requirement will be no less than 50tb.

$$\frac{637\times360\times60\times10\times30\times9(Byte)}{1024\times1024\times1024}\approx2TB/YEAR \qquad (5\text{-}1)$$

3) SAMPLING FREQUENCY

The sampling frequency is also called the sampling speed or sampling rate. When agricultural engineering machinery is performing agricultural operations, information such as motion posture, component revolutions, temperature, etc., continuously changes over time. That is to say, only when sampling a myriad of data per unit time, it is possible to restore the change of the entire motion posture with almost no error [27]. However, there are three factors that restrict it.

First, according to physics Equation 5-2 (frequency versus periodic relationship), where f represents frequency and T represents period. The infinite number of samples per unit time means that the single sampling period is infinitely small, but the current technical level does not meet the requirements.

Second, the infinite number of samples per unit of time means that the temporary space required to store data is infinite. This is not possible for in-vehicle mobile data centers and lower-level machines with temporary space of less than 100G.

Third, even though the above two limitations have been compromised, the central processing unit of the in-vehicle mobile data center and the lower computer is far from enough to process an infinite amount of data. Because the data sampled by the motion tracking microprocessor is placed first into the internal registers, the internal registers are a one-way queue that satisfies the FIFO (first incoming data leaving) rule. When the external read is much slower than the speed of the motion tracking microprocessor and the internal registers are fully loaded, the motion tracking microprocessor will discard the data being acquired. New data will be written when there is data in the queue.

$$f=\frac{1}{T} \qquad (5\text{-}2)$$

According to the actual situation, in the more mature acquisition schemes currently available, the sampling frequency of the motion tracking microprocessor with higher precision is about 1 $\sim$ 100KHz. According to the actual situation, in the more mature acquisition schemes that are currently available, the sampling frequency of the motion tracking microprocessor with higher precision is about 1 $\sim$ 100KHz. With higher precision lower position machine GPIO (general input / input bus expansion) clocked at about 10 $\sim$ 20MHz, the lower computer does not cause the bottleneck of the sensor sampling frequency. However, when the lower computer or the vehicle mobile data center simultaneously accesses multiple sensor sampling data, at the processing level multiple sensors need to be serially processed according to the hard-coded logic of the handler, so that the calculated main frequency of the lower-level machine GPIO actually loaded on each sensor is greatly reduced. According to the five-five principle of computer system design performance indicators, when the demand performance parameters are determined, it is necessary to preserve the same performance redundancy as the actual demand in case of subsequent software and hardware upgrades, and to ensure system stability when the system fluctuates. If the state of lower machine GPIO frequency 10MH is followed, then the actual sensor sampling frequency after all serials is 5MHz. According to each remote data center accessing 100 in-vehicle mobile data centers, each in-vehicle mobile data center simultaneously accesses the estimation of 100 data collection points, and the sampling frequency of 100 sensors is 5MHz. The average sampling frequency for each sensor is therefore approximately 50KHz, which just meets the sampling frequency of a high precision microprocessor sensor [28].

4) MINIMUM UNIT AND ERROR

The minimum unit is the minimum effective difference of the range of the parameter variable selectable range. In the sensor acquisition, because the sensor data bus bit width is larger than that of the general-purpose computer, it cannot be stored and operated by floating point values. And because most sensors have many different ranges, they can be manually set by the user to adapt to different scenes, thus introducing the concept of error. In the data acquisition subsystem of the management analysis platform, the sensor is hard-wired to the lower computer through a direct connection. Since the sensor's register data bus width is low, it is usually necessary to divide the value of one measurement into different data segments and store them in different registers. When the lower computer reads a complete measurement, it needs to read different registers multiple times, and then integrate the

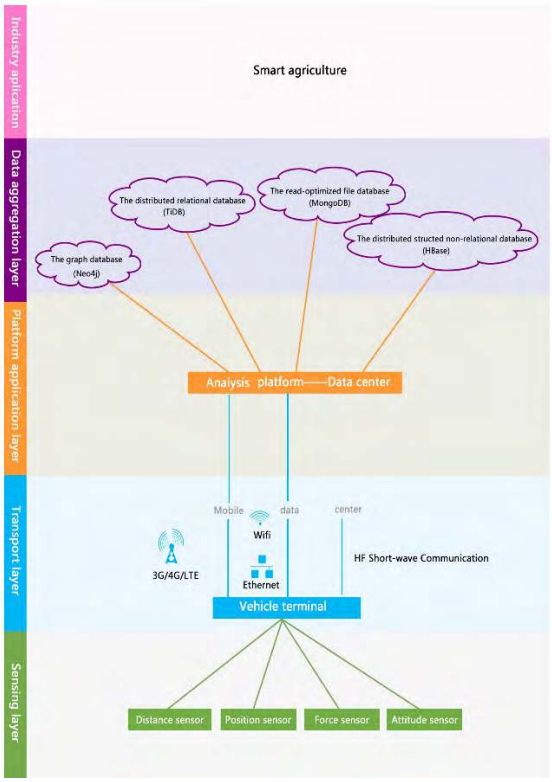

FIGURE 5. Internet of things architecture diagram.

data according to the agreed data segmentation order to get the final measurement value.

The design characteristics of the sensor board is limited by the use of metal wiring. When the metal wire passes the high-frequency signal, due to the Faraday electromagnetic effect, it will inevitably cause fluctuations in the signal transmitted between the lower computer and the sensor or between the register and the micro-processing chip, causing errors. At this time it is necessary to make a reasonable range selection based on the sensor's percentage error ratio to minimize the theoretical error, and a software algorithm is needed to correct the error [29].

In the data acquisition system, when the data is collected at a data point within 200 times (including 200 times) of the minimum unit, the percentage error ratio is required to be less than 1%. When the range is less than 200 times the minimum unit and within 500 times (including 500 times), the percentage error ratio is required to be less than 3%. When the range is more than 500 times the minimum unit, the percentage error ratio is required to be less than 5%.

B. SYSTEM ARCHITECTURE

The Internet of Things architecture diagram of this system is shown in Figure 5.

The overall architecture model of the management analysis platform is shown in Figure 6: the data acquisition subsystem, the data storage subsystem, the data analysis subsystem, and

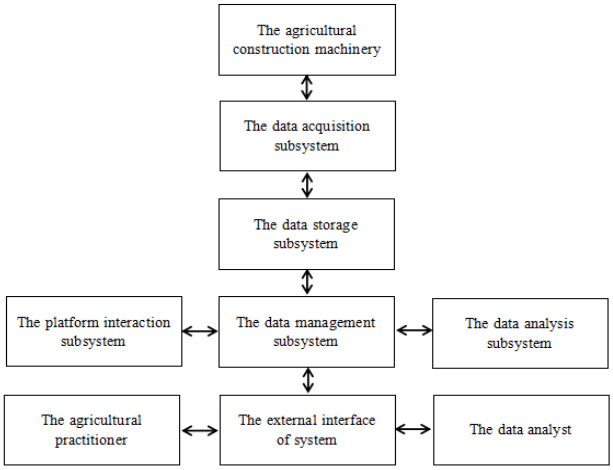

FIGURE 6. Overall architecture model diagram.

the platform interaction subsystem work around the data management subsystem. There are four input and output of the management analysis platform: the data acquisition of the sensor data collected by the data acquisition subsystem, the input of the analysis model of the data analyst in the data analysis subsystem, and the agricultural practitioner of the platform interaction subsystem. There is also an external interface reserved for data transmission with third party systems in the data management subsystem.

1) DATA ACQUISITION SUBSYSTEM

As the management analysis platform, the data acquisition subsystem is the only subsystem with a single input. The architecture model is shown in Figure 7. The raw data collected by the four types of sensors is received, integrated, packaged, and transmitted by the lower computer. Then the data sends to the car mobile data processing center. After data screening and analysis, the vehicle mobile data processing center determines whether it is necessary to modify the configuration parameters of each sensor to adapt to the real-time operating conditions of the data acquisition and reduce the error. If the configuration needs to be modified, the control command of the target configuration is calculated and sent to the sensor that needs to change the configuration via the host computer. If it is not necessary to modify the configuration information of the sensor, the data file integrated and packaged by the lower computer is sent to the data management subsystem through the WLAN network or the mobile transmission network.

When the data acquisition subsystem performs dynamic adjustment of self-feedback data acquisition, the main processing logic is shown in Figure 8. When the vehicle mobile data processing center detects the data collected by the sensor currently received by the lower computer, when some of the continuous time series data triggers the dynamic adjustment data threshold, the real-time calculation is performed according to the preset parameters of the sensor, and data

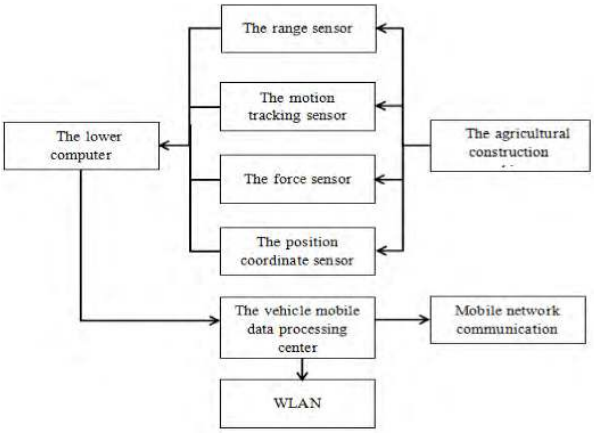

FIGURE 7. Data acquisition subsystem architecture model diagram.

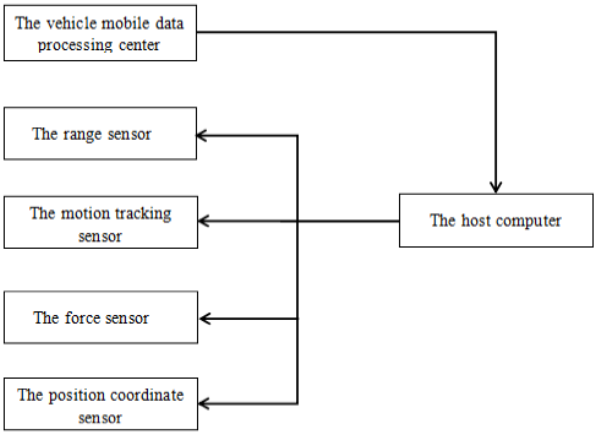

FIGURE 8. Data acquisition dynamic adjustment model diagram.

acquisition logic re-planning is performed. The logic is converted into a control signal and sent to the lower computer. The lower computer sends a parameter adjustment command to the corresponding sensor according to the physical address of the sensor and a signal line directly connected to the sensor, achieving the purpose of dynamic adjustment of data acquisition.

2) DATA STORAGE SUBSYSTEM

The data storage subsystem records all the data of the management analysis platform. The architecture model is shown in Figure 9: The IO load balancer is the only data entry and exit path of the data storage subsystem. All data write and read requests must be processed by the IO load balancer. After the IO load balancer authenticates the requested authority, it forwards the request to the data adapter according to the system's load pressure and data requirements. The data adapter writes data written to the type operation request to the data buffer pool and notifies the data scheduler to process it. For requests that read type operations, the data adapter is directly notified to the data scheduler operation after marking [30].

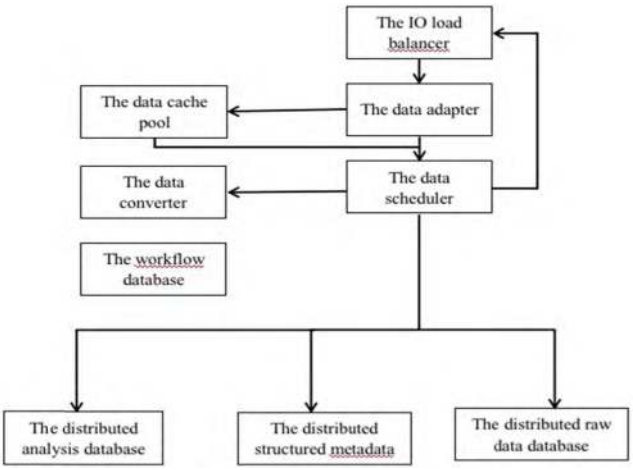

FIGURE 9. Data storage system architecture model diagram.

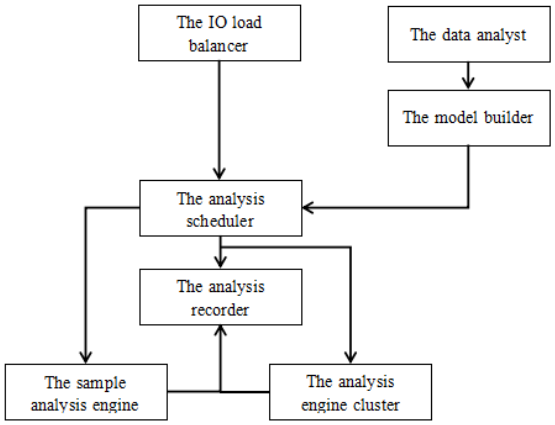

FIGURE 10. Data analysis subsystem architecture model diagram.

The last step in the data scheduler's execution of the actual store, is to forward to a different database based on the data type. Working conditions raw data, structured metadata, and analytical data are directly forwarded to the corresponding database cluster for reading and writing. For the writing of workflow data, it needs to be converted into a database-friendly storage type and stored through the workflow graph conversion algorithm. The read operation is the reverse operation of the write operation.

3) DATA ANALYSIS SUBSYSTEM

The data analysis subsystem is the analysis engine of the management analysis platform. The architecture model is shown in Figure 10.

- The IO load balancer functions the same as the IO load balancer of the data storage subsystem.
- Through the model builder, the data analyst forms an abstract analysis algorithm model and theory to form an analysis process work list (a working directory table) that the analysis engine can actually refer to. The analysis scheduler matches the working directory table according to the trigger mechanism, the scheduler and the sample size, and respectively schedules the sample analysis engine and the analysis engine cluster.

• The analysis recorder records all of the analyzed processes and intermediate data sets so that the data analyst can modify the analytical algorithm model. And the analysis of the data recorded by the recorder can also be used for analysis playback and breakpoints.

C. DATA STORAGE

The data storage subsystem mainly completes the storage of the original data of the working conditions, the storage and retrieval of the structured metadata, the storage and modification of the model data, and the storage of workflow of the management analysis platform.

The storage of the raw data of the working condition refers to the data generated by the raw data of the agricultural engineering machinery collected by the sensor through the processing of the data acquisition subsystem for storage. The data collected by the data acquisition subsystem will be in both text format and binary format.

The storage and retrieval of structured metadata refers to the storage of data after data structure processing of the raw data of the text format and the reading of data with retrieval. When the analysis platform performs model analysis, structured metadata is directly called and needs to be filtered according to conditions such as time, structure partitioning, and data balance.

The storage and modification of the model data refers to the basic data of the analysis model after processing of the working condition data. Most of this data comes from the input of the model analyst, and some of the data is intermediate data of the serial sub-analysis process performed serially.

The storage of the workflow of the management analysis platform refers to business logic, the storage, modification, reading of data, and processes necessary for the management analysis platform as a complete business system.

According to all the requirements of the platform storage subsystem mentioned above, it is summarized that there are five types of databases: distributed data warehouse, distributed relational database, read optimized file database, graphic database, and distributed structured non-relational database.

The primary purpose of a distributed data warehouse in a data storage subsystem is to store raw data. After the raw data of the working condition is generated by the sensor, it will not be deleted in the management analysis platform, and the data will not be directly read and written by other subsystems in the platform. Moreover, the read performance of this data is not high. Hive is selected as the data engine of the distributed data warehouse, which performs a large amount of file storage and does not require the stored data to have a clear format and specification.

The primary purpose of a distributed relational database in a data storage subsystem is to store structured analysis metadata and structured case data. Since the working condition data is not deleted and has the structural characteristics at the same time, TiDB is selected as the distributed relational database. Because TiDB can be highly compatible with RDBMS of relational database, and can support unlimited horizontal expansion of MySQL, it solves the problem that the large amount of metadata after structuring leads to the failure of traditional relational database storage [31].

The primary purpose of the read-optimized file database in the data storage subsystem is to store temporary structured data of the original case data. Because the format of the structural chemical data required for the analysis process may have changed in the case of analysis of model and algorithm changes, the data format determined by the new model and algorithm cannot be determined. In order to ensure that the new analysis models and algorithms can be used normally, and the rollback of the system will not be affected after new analysis models and algorithms fail, the new analytical models and the structured data required by the algorithms are stored separately. According to the demand scenario of reading less writes and data format is not fixed, MongoDB is selected as the data engine of the read optimized file database [32].

The main function of the graph database in the data storage subsystem is to store the visualization files into which the data analysis results are converted. Each time through the platform interaction, subsystem query analysis results must be re-formatted and provide file export function. However, the conversion of the analysis results into a page visualization file takes a long time, resulting in poor user experience. This paper proposes a method to directly convert the stored analysis results into a visualization file, which not only shortens the response time, but also facilitates file export. Because the visualization file has the characteristics of unfixed format and complicated association, a single text or relational database does not meet the requirements and so the Neo4j graphical database is used as the data engine [33].

The distributed structured non-relational database in the data storage subsystem is the most complex and content-changing database in the platform. The main role is to record the data analysis process and intermediate analysis results. Because data analysis is a process of large sample data and large-scale batch processing, not only does the database that stores data have high scalability, but also requires high read and write performance. Based on this, HBase is adopted as the distributed structured non-relational database engine [34]. Like the distributed data warehouse Hive, HBase's scalability is also based on the Hadoop platform. However, compared to Hive, HBase has a high carrying capacity in a clear analysis subject scenario, which can provide high-speed data writing and data reading [35].

D. DATA ANALYSIS

The data analysis subsystem in this study is similar to the data analysis platform, which refers to the computer software system that actually runs the data analysis algorithm and analysis model and can output the analysis results. In this study, the MapReduce data analysis engine based on Hadoop technology and the Spark engine compatible with Hadoop technology are mainly used.

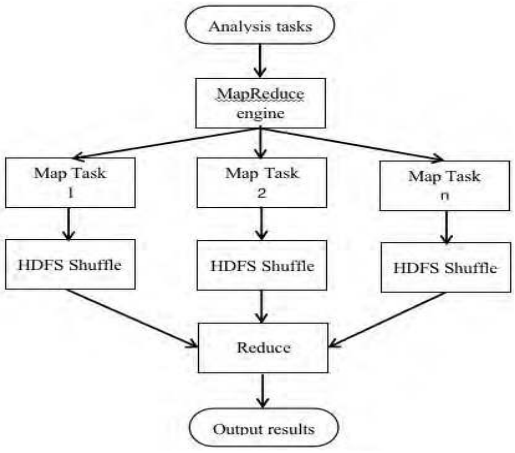

FIGURE 11. MapReduce task flow chart.

As shown in Figure 11, the MapReduce data analysis engine solves the problem that a single analysis host cannot save all data through the HDFS cluster and the easy-to-expand feature of Hadoop. The analysis task is split into multiple jobs and submitted to multiple hosts in the cluster for analysis, which solves the problem of a single host not being able to handle complex tasks. This feature of MapReduce is exactly the same as the value analysis task for large samples and super samples in the data analysis subsystem. Each sub-task of MapReduce will write the result on the HDFS disk and can conveniently record the intermediate process of washing. So MapReduce is used as part of the data analysis subsystem to build the analysis engine cluster [36].

While the MapReducce data analysis engine writes the intermediate results of the analysis process to the HDFS disk, it is inevitable that it will be subject to the speed of the disk IO (input and output), resulting in a longer time spent in the analysis process. This fails to guarantee a strong real-time requirement for the fuse type and threshold type analysis tasks that the data analysis subsystem needs to process in near real time. For this, the Spark data analysis engine is required [37].

The Spark data analysis engine can quickly process data analysis tasks. It has more than ten times faster performance than the MapReduce data engine when processing single-read, convert, and load-type analysis tasks. The memory processing is fast, but when the amount of data to be analyzed is larger than the memory capacity, data analysis cannot be performed. If the data analysis is forced, the analysis results will be invalid or the analysis will be aborted, and the entire analysis engine software will be partially unusable. So in this study, the Spark data engine can only be used as a sample analysis engine [38].

VI. EXPERIMENTS AND RESULTS

A. EXPERIMENTS PROCESS

1) SYSTEM INITIAL DATA PREPARATION

The initial data of the system is derived from the 13th division of Hongshan Farm of Xinjiang Production and Construction Corporation of China, the Dashan Daxueshan Agricultural Machinery Service Professional Cooperative of Qinghai Province, and the Meike Soybean Planting Farmers Professional Cooperative of Xunke County, Heilongjiang Province. The data covers 60 sets of agricultural engineering machinery of different models and different uses, such as Dongfanghong LX904, Dongfanghong LX1204 and John Deere 6J-1854. The data such as model, physical size, number of parts and model of these machines must be input into the system, especially the modeling of physical data, which is of great significance for analyzing the cause of the fault and providing references for maintenance. In this design, 3D Engineering and Simulator software were used to perform physical modeling based on the data information of agricultural machinery parts provided by General Electric Company. The following describes the physical model introduction of the centrifugal governor as an example.

TABLE 2. Modeling sequence table.

Step number	Specific operation
1	The spindle shaft and the shaft key are selected, and the shaft key is placed in a predetermined shaft key groove of the spindle shaft, and the shoe fork rod is selected according to one-half length of the distance between the two shaft keys at the end.
2	The fork bushing and the arm bushing are arranged in the order of the spindle shaft and fixed by the shaft key, and the shoe fork lever is connected to the lower part of the fork bushing through the positioning post.
3	The gravity arm, the force arm fork rod and the positioning rod are combined and connected into the sleeve, wherein the end of the force arm is connected with the arm sleeve, and the end of the force arm fork is introduced at the upper end of the shoe sleeve.
4	Connect the float ball to the lower end of the gravity arm, connect the friction shoe to the end of the shoe fork, and fix the helical gear to the lower end of the spindle shaft through the shaft key.
5	Place the friction shoe inside the tympanic casing and combine it with the software combination button.
6	According to the combination prompt in the previous step, after assembling in reverse order, click Save.

The physical modeling of the watt-type centrifugal governor is performed by the 3D Engineering software according to the steps of the modeling sequence table in Table 2. The process is shown in the model diagram of the centrifugal governor in Figure 12.

2) PARAMETER ZEROING

Due to the tolerance of the process level, the micro-processing chip of the sensor will inevitably have an initial error in production. At the same time, because most of the sensors in this experiment are hand-welded and installed, when the

FIGURE 12. Centrifugal governor model diagram.

TABLE 3. Attitude data table.

Roll	Pitch	Yaw	Temp
2.68°	0.68°	1.49°	28.67°C
2.64°	0.63°	1.49°	28.68°C
2.60°	0.59°	1.49°	28.68°C
2.57°	0.54°	1.49°	28.68°C
2.54°	0.50°	1.49°	28.68°C
2.50°	0.46°	1.50°	28.68°C
2.47°	0.41°	1.51°	28.68°C
2.44°	0.37°	1.52°	28.68°C
2.41°	0.32°	1.53°	28.68°C
2.38°	0.28°	1.54°	28.68°C

sensor is fixed the secondary error outside the measurement will be caused by the distortion of the pin and the deformation of the expansion plate. Therefore, before the motion attitude sensor officially collects data, it is necessary to zero the sensor's posture so that the agricultural machinery item is in a static state and the data collected by the sensor is the design zero error of the current sensor collection point. After obtaining the zero-error value, the original data is subjected to zero-error compensation, and more accurate sampling data is obtained. Table 3, the attitude data table, shows the pose data for ten consecutive data points.

For the data collected by the position coordinate sensor, the information of each data point is shown in Table 4.

According to data from the satellite signal data table, the ephemeris data of the corresponding satellite is queried in the databases of GPS, Glonass, Beidou navigation, and Galileo. According to the ephemeris data, the relative position coordinates of each satellite are obtained. The ephemeris map after coordinate fusion is shown in Figure 13.

After calculation by the relevant calculation method in the foregoing, the data coordinate type WGS-84 is obtained, the longitude is 126.626864°, the latitude is 45.714131°, and the precision is 0.37 meters.

TABLE 4. Satellite signal data sheet.

Azimuth	Height angle	Type	SVID
251°	2°	GPS	3
317°	16°	GPS	9
167°	11°	GPS	14
262°	61°	GPS	16
80°	27°	Glonass	10
295°	31°	Glonass	2
168°	7°	Glonass	22
299°	35°	Glonass	12
5°	65°	Glonass	11
229°	18°	Glonass	1
45°	34°	Glonass	20
120°	44°	Glonass	21
343°	7°	Glonass	3
160°	35°	Beidou	1
230°	21°	Beidou	2
201°	34°	Beidou	3
136°	27°	Beidou	4
251°	4°	Beidou	5
195°	30°	Beidou	6
286°	73°	Beidou	7
219°	46°	Beidou	9
270°	48°	Beidou	10
168°	28°	Beidou	12
224°	8°	Beidou	14
234°	34°	Galileo	4
185°	4°	Galileo	9

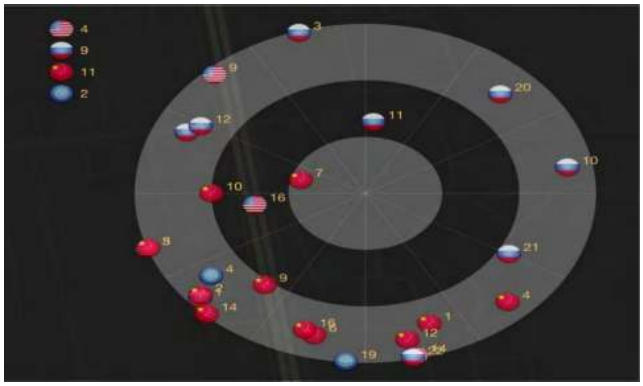

FIGURE 13. Ephemeris.

3) EXPERIMENTAL SAMPLE COLLECTION

After the preparation of various experiments in the system, the data acquisition equipment, such as sensors, were installed on 60 agricultural engineering machines. The floor data was then collected from September 10, 2018 to October 8, 2018 without additional human intervention. During the period, the total number of working hours of agricultural engineering machinery in the field operation was represented by over 168 million data points and 1.8 tb of storage space.

B. SYSTEM INTERFACE

After the completion of the overall system of the agricultural machinery failure analysis platform based on big data, it consisted of six menu items: system overview, data playback, data import, workflow analysis, device management, and system setup.

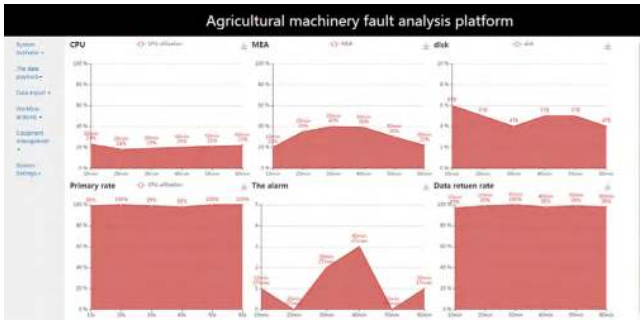

FIGURE 14. System overview.

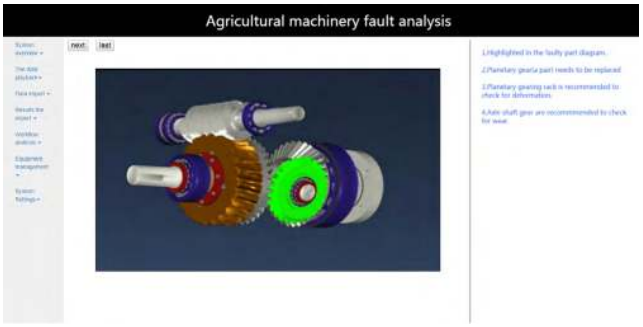

FIGURE 15. Model alarm map.

The system overview page is shown in Figure 14. The left side of the page displays the action menu that can be clicked in the form of a list, and the right area shows the monitoring panel of the current system. Through the monitoring panel, you can intuitively view the basic information of each important indicator of the current fault analysis platform software system, such as CPU usage, memory usage, and data footprint on the disk. In addition, it also displays the ratio of one analysis and calculation of the raw data of the working condition, highlights the current system alarm quantity, and has real-time data return rate monitoring, which can intuitively obtain the load status of each vehicle's mobile data center.

C. EXPERIMENTAL RESULTS

In the 60 sets of agricultural engineering machinery collected in the experimental samples, when the anomaly analysis model was analyzed it was found that there were 13 abnormal mechanical rear axle power output shaft parameters. This was mainly manifested in the fact that the motion attitude data irregularly appeared to be unstable in the axis of the rotation path, and the rotational speed parameter had a nonlinear jump. After inspection by the farm machinery maintenance personnel and experimental personnel, it was found that 13 machines did not show obvious abnormalities, and the sensors installed on the power output shaft of the rear axle of the two machines (hereinafter referred to as A and B cars) were loose. Since the farm machinery maintenance personnel did not find any obvious abnormalities, the sensor was re-fixed and put into use. In the subsequent 100 working hours, the A and B vehicles broke the power output shaft.

According to the fault data of A car and B car, the Alstom company's s8 series model was disassembled, merged and reconstructed with China's Zhongqing Qing's ds9x model, and the fuse data adjusted by adjusting the vehicle mobile data center. In the 20 working hours after the A and B vehicles were repaired and replaced with the power output shaft, the same position sensor reported the data abnormality again. The fuse-type data analysis report of the other four machines (hereinafter referred to as C car, D car, E car, F car) after fault adjustment showed that the power output shaft parameters of the four machines reached the fault critical value. After the motion posture sensor was attached to the auxiliary mounting equipment (power output shaft receiving end) of the C, D, E, and F vehicles, the four vehicles continued to perform farm work. In the 10 working hours after the restart of the four cars, the four vehicles successively broke the power output shaft. At this time, the farm administrator requested that all 13 agricultural engineering machinery with the same alarm stop production and wait for the complete analysis report of the management analysis platform.

The data analyst modeled and analyzed the fault critical data of the six faulty machines and the data of the subsequent four mechanical attached equipment. It was found that the direct cause of the power output shaft fracture was the planet deformation carrier in the differential of the power output shaft.

TABLE 5. Event record table.

Serial number	Working record point	vehicle	event
1	302	All 13 units	Modification of auxiliary mounting equipment, replacement of output universal joints
2	343	A car	First appearance of power output shaft data abnormal alarm
3	396	B car	The output shaft is broken and the data is normal after replacement.
4	435	A car	The output shaft is broken and the data is normal after replacement. Also found that the output shaft oil seal seepage
5	437	All 13 units	Adjust the onboard data center fuse analysis parameters
6	451	C car	First failure threshold warning
7	452	CDEF four cars	Adding attached mounting equipment motion attitude sensor
8	459	F car	Output shaft break
9	468	All 13 units	Stop the operation and return to the factory to wait for the test result

TABLE 6. Alarm statistics table.

Classification	Parts	Data point
Total	713	16.8 billion
Total number of alarms	120,000	827,000
Direct associations	60,000	468,000
Direct association ratio	50.0%	56.6%
Indirect associations	38,000	210,000
Indirect association ratio	31.7%	25.4%
Effective associations	98,000	678,000
Effective association ratio	81.7%	82.0%
Invalid link count	22,000	149,000
Invalid association ratio	18.3%	18.0%

Figure 15 shows the alarms obtained when the management analysis platform performed the third abnormal analysis. The alarm can be accurate to the component level.

The above faults are arranged in time sequence as shown in Table 5.

During the trial, the recorded data points totaled 1.68 billion data points, and the total number of sampled parts was 713. The specific alarms and associations are shown in Table 6. Direct correlation refers to the direct causal relationship between the alarmed information and the component or data of the final problem. Indirect association means that the information being alerted has an indirect causal relationship with the component or data of the final problem.

It can be seen from the table that in this study, the effective correlation of the component alarms and data point alarms of the 60 sets of agricultural engineering machinery involved has exceeded 80%. This shows that at the level of accuracy of the analysis, the management analysis platform has practical value by providing early warnings and reducing failures. A complete fault record and analysis report can also be given after the fault has occurred.

VII. CONCLUSION

This study popularized the basic knowledge of agricultural big data through the introduction of international research on big data and smart agricultural machinery. At the same time, this study discovered that China's current research in the field of big data analysis of agricultural machinery is relatively scarce, expounding the importance and necessity of the intelligent application of agricultural engineering machinery.

This paper demonstrates the practical significance of big data analysis technology in agricultural production, especially in agricultural engineering machinery maintenance and fault detection, through the introduction of the whole process from the beginning of data collection to the output of big data analysis results. It also verifies the feasibility of analyzing agricultural engineering machinery through big data technology. Through the organic integration of multiple algorithms and multiple models, the platform enables fast-blown data analysis and data analysis of big data oversamples according to specific needs.

In the experiment, the effective correlation of component alarms and data point alarms of agricultural engineering machinery exceeded 80%. The complete and meticulous model enables those who do not have computer domain expertise and mechanical maintenance knowledge to intuitively understand analytical reports and reduce the experience of mechanical maintenance.

Overall, the overall study makes a significant contribution to knowledge in the field of digital and smart agriculture.

REFERENCES

[1] M. Mijiti, P. Silamu, and M. Tuohuti, "Analysis and reflections on some problems of agricultural mechanization," *Social Sci., Educ. Hum. Sci.*, pp. 16–18, 2017.
[2] I. M. Carbonell, "The ethics of big data in big agriculture," *Internet Policy Rev.*, vol. 5, no. 1, pp. 23–25, 2016.
[3] S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, "Big data in smart farming—A review," *Agricult. Syst.*, vol. 153, pp. 69–80, May 2017.
[4] K. L. Krishna, O. Silver, W. F. Malende, and K. Anuradha, "Internet of Things application for implementation of smart agriculture system," in *Proc. IEEE Int. Conf. I-SMAC (IoT Social, Mobile, Anal. Cloud) (I-SMAC)*, Feb. 2017, pp. 54–59.
[5] G. B. Horn, M. Griot, and R. Subramanian, "Method and apparatus for managing packet data network connectivity," Google Patents 9 386 607 B2, Jul. 5, 2016.
[6] Y. Kushita, M. Iwamoto, K. Hasegawa, and S. Sonoda, "Tractor," Google Patents, 2016.
[7] O. S. Mensah, "Chinese agricultural sector: A review of prospects and challenges," vol. 11, no. 4, pp. 1–12, 2017.
[8] F. Meng, D. Yang, Y. Wang, and Y. Zhang, "Study of agricultural machinery operating system based on Beidou satellite navigation system," in *Proc. Int. Conf. Intell. Interact. Syst. Appl.* Cham, Switzerland: Springer, 2017, pp. 268–273.
[9] M. Schneider, "What, then, is a Chinese peasant? *Nongmin* discourses and agroindustrialization in contemporary China," *Agricult. Hum. Values*, vol. 32, no. 2, pp. 331–346, 2015.
[10] A. Weersink, E. Fraser, D. Pannell, E. Duncan, and S. Rotz, "Opportunities and challenges for Big Data in agricultural and environmental analysis," *Annu. Rev. Resource Econ.*, vol. 10, pp. 19–37, Oct. 2018.
[11] Q. Zhou, D. Zou, and P. Liu, "Hybrid obstacle avoidance system with vision and ultrasonic sensors for multi-rotor MAVs," *Ind. Robot, Int. J.*, vol. 45, no. 2, pp. 227–236, 2018.
[12] F. Jachmann, M. Engler, B. Rothenberger, and T. Dollmann, "Distance-measuring sensor and method for detecting and determining the distance of objects," Google Patents 9 995 820 B2, Jun. 12, 2018.
[13] A. Leick, L. Rapoport, and D. Tatarnikov, *GPS Satellite Surveying*. Hoboken, NJ, USA: Wiley, 2015.
[14] A. Perrig, R. Szewczyk, J. D. Tygar, V. Wen, and D. E. Culler, "SPINS: Security protocols for sensor networks," *Wireless Netw.*, vol. 8, no. 5, pp. 521–534, Sep. 2002.
[15] W. Li and J. Wang, "Effective adaptive Kalman filter for MEMS-IMU/magnetometers integrated attitude and heading reference systems," *J. Navigat.*, vol. 66, no. 1, pp. 99–113, 2013.
[16] B.-H. Wang, C.-Y. Zhao, and Y.-Y. Liu, "An improved SAR interferogram denoising method based on principal component analysis and the Goldstein filter," *Remote Sens. Lett.*, vol. 9, no. 1, pp. 81–90, 2018.
[17] R. V. Garcia, H. K. Kuga, and M. C. F. P. S. Zanardi, "Unscented Kalman filter applied to the spacecraft attitude estimation with euler angles," *Math. Problems Eng.*, vol. 2012, Sep. 2012, Art. no. 985429.
[18] M. Wang, "City data fusion: Sensor data fusion in the Internet of Things," *Int. J. Distrib. Syst. Technol.*, vol. 7, no. 1, pp. 15–36, 2016.
[19] X. Wang, R. Bai, X. Cui, T. Wu, and Z. Qian, "Research on data fusion algorithm for attitude detection systems based on MEMS and magnetoresistive sensors," in *Proc. 9th Int. Conf. Adv. Infocomm Technol. (ICAIT)*, Nov. 2017, pp. 68–74.

[20] Z. Zhu, "Change detection using Landsat time series: A review of frequencies, preprocessing, algorithms, and applications," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 370–384, Aug. 2017.
[21] E. Malikov and Y. Sun, "Semiparametric estimation and testing of smooth coefficient spatial autoregressive models," *J. Econ.*, vol. 199, no. 1, pp. 12–34, 2017.
[22] C. Lizama, "The Poisson distribution, abstract fractional difference equations, and stability," *Proc. Amer. Math. Soc.*, vol. 145, no. 9, pp. 3809–3827, 2017.
[23] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long short term memory networks for anomaly detection in time series," in *Proceedings*. Ottignies-Louvain-la-Neuve, Belgium: Presses Universitaires Louvain, 2015, p. 89.
[24] J. Huang, Q. Zhu, L. Yang, and J. Feng, "A non-parameter outlier detection algorithm based on natural neighbor," *Knowl.-Based Syst.*, vol. 92, pp. 71–77, Jan. 2016.
[25] D. C. Wickramarachchi, B. L. Robertson, M. Reale, C. J. Price, and J. Brown, "HHCART: An oblique decision tree," *Comput. Statist. Data Anal.*, vol. 96, pp. 12–23, Aug. 2016.
[26] L. Barton, B. Wolf, D. Rowlings, C. Scheer, R. Kiese, P. Grace, K. Stefanova, and K. Butterbach-Bahl, "Sampling frequency affects estimates of annual nitrous oxide fluxes," *Sci. Rep.*, vol. 5, Nov. 2015, Art. no. 15912.
[27] Y.-H. You, J. B. Kim, and H.-K. Song, "Pilot-assisted fine frequency synchronization for OFDM-based DVB receivers," *IEEE Trans. Broadcast.*, vol. 55, no. 3, pp. 674–678, Sep. 2009.
[28] S. Tu and L. Xu, "A theoretical investigation of several model selection criteria for dimensionality reduction," *Pattern Recognit. Lett.*, vol. 33, no. 9, pp. 1117–1126, 2012.
[29] T. Sun, Y. Song, Y. Li, and L. Xu, "Separation of comprehensive geometrical errors of a 3-DOF parallel manipulator based on Jacobian matrix and its sensitivity analysis with Monte-Carlo method," *Chin. J. Mech. Eng.-English Ed.*, vol. 24, no. 3, p. 406, 2011.
[30] A. Gulati, I. Ahmad, and C. Kumar, "Distributed storage resource scheduler and load balancer," Google Patents, 2014.
[31] Y. Wu and H. Guan, "Biggy: An implementation of unified framework for big data management system," 2018, *arXiv:1810.09378*. [Online]. Available: https://arxiv.org/abs/1810.09378
[32] S. Chickerur, A. Goudar, and A. Kinnerkar, "Comparison of relational database with document-oriented database (MongoDB) for big data applications," in *Proc. IEEE 8th Int. Conf. Adv. Softw. Eng. Its Appl. (ASEA)*, Nov. 2015, pp. 41–47.
[33] C. Vicknair, M. Macias, Z. Zhao, X. Nan, Y. Chen, and D. Wilkins, "A comparison of a graph database and a relational database: A data provenance perspective," in *Proc. ACM 48th Annu. Southeast Regional Conf.*, 2010, p. 42.
[34] P. Gong, C. Lv, Y. Gong, H. Ma, Y. Sun, and L. Wang, "Research on keyword retrieval method of HBase database based on index structure," *AIP Conf. Proc.*, vol. 1890, no. 1, 2017, Art. no. 040064.
[35] K. Kolonko, "Performance comparison of the most popular relational and non-relational database management systems," Tech. Rep., 2018.
[36] J. Morán, C. de la Riva, and J. Tuya, "Testing MapReduce programs: A systematic mapping study," *J. Softw., Evol. Process*, vol. 31, no. 3, p. e2120, 2019.
[37] A. Eldawy and M. F. Mokbel, "SpatialHadoop: A mapreduce framework for spatial data," in *Proc. IEEE 31st Int. Conf. Data Eng.*, Apr. 2015, pp. 1352–1363.
[38] M. Armbrust, R. S. Xin, C. Lian, Y. Huai, D. Liu, J. K. Bradley, X. Meng, T. Kaftan, M. J. Franklin, A. Ghodsi, and M. Zaharia, "Spark SQL: Relational data processing in spark," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2015, pp. 1383–1394.

DAN LI received the Doctorate degree in engineering. He is currently the Director of computer science and technology with Northeast Forestry University. He is also a Visiting Scholar with the University of Georgia. He has published more than 20 papers on SCI and EI papers. He is a member of the Computer Branch of the China Forestry Society and a member of the National Forestry and Grassland Internet of Things and Artificial Intelligence Application Technology Innovation Alliance. He also presided over seven projects, including the National Spark Program, the National Forestry Administration's Public Welfare Industry Project, Heilongjiang Province Science and Technology Research, and the Heilongjiang Provincial Natural Science Foundation. He served as a Reviewer for several journals, such as the *Journal of Beijing Forestry University* and the *Journal of Northeast Forestry University*.

YI ZHENG, photograph and biography not available at the time of publication.

WEI ZHAO, photograph and biography not available at the time of publication.