Eng Int Syst (2024) 5: 507–520

Engineering
Intelligent Systems

Intelligent Security System in Food Supply Chain Based on Big Data Analysis

Xin Zhang*

College of Business, Jiaxing University, Jiaxing 314001, Zhejiang, China

Traditional security management systems established and applied in the food supply chain mostly use relational databases and parallel data warehouses to store, manage and analyze data. With the continuous development of technology, today's data sources are more diverse, and traditional methods no longer meet the requirements for data processing efficiency and availability. Therefore, big data analysis can be used to establish an intelligent security system in the food supply chain, which segments massive data, decomposes tasks, and summarizes results. This facilitates intelligent security monitoring and the management of every link in the food supply chain specialization on food supply chain platform. Firstly, a large amount of security-related data in the food supply chain is collected through Internet of Things (IoT) devices and radio frequency identification (RFID) technology, including sensor data, monitoring data, etc. Then, the Hadoop Distributed File System (HDFS) is selected to store the collected raw data. At the same time, the data is preprocessed for subsequent analysis and processing. The preprocessing involves data cleaning, deduplication, conversion, and standardization. Various big data analysis techniques such as Spark and k-means algorithms are utilized to analyze and mine stored data in order to detect abnormal patterns, identify risk events and behaviors, and predict potential threats. Data visualization tools are used to present the analysis results in a way that is easy to understand and visualize. Displaying sensor data, food source information, etc., can help users understand security trends and key indicators in real time. Based on the analysis results, some security responses and decision-making processes are automated by, for instance, automatically invoking security measures and emergency response processes to quickly handle and mitigate security incidents. By continuously optimizing and analyzing models, algorithms, and strategies, as well as learning and training new data, the accuracy and response efficiency of intelligent security systems can be improved. In this study, when using big data analysis technology to process data, the average response time for reading data is 3.647 seconds, and the average response time for writing data is 6.139 seconds, with an average throughput of 896MB/s. Compared with traditional methods, this is a significant improvement of data processing speed, reflecting its stronger parallel processing ability. Big data analysis has good scalability in data processing, and can therefore handle a mixture of multiple types of data. It can also comprehensively and efficiently control security throughout the entire platform for food supply chain specialization to build an intelligent security system. The application of intelligent security systems based on big data analysis in the food industry chain can improve the safety, quality, and traceability of food.

Keywords: Internet of Things Devices, Hadoop Distributed File System, Intelligent Security System, Food Supply chain, Radio Frequency Identification

1. INTRODUCTION

In the modern economy, the platform of a food supply chain has become one of the main features of the food industry [1]. Compared with traditional food supply chains, platform food supply chain platform not only covers production, processing, transportation and other links, but also incorporates emerging technologies such as e-commerce platforms and logistics platforms, digitizing and networking the entire supply chain process, improving efficiency and visibility. However, with the rapid development of food supply chain platform, more challenges have emerged, particularly in relation to food safety management. Establishing an intelligent security system for the food supply chain can prevent food safety issues, achieve full monitoring and traceability of the food supply chain, and strengthen data analysis and decision support. Compared with traditional methods, intelligent

*Address for correspondence: Xin Zhang, College of Business, Jiaxing University, Jiaxing 314001, Zhejiang, China, Email: zcxin9@163.com

security systems based on big data analysis can collect data from multiple stages of the food supply chain and collect more complex and large-scale data, as well as provide comprehensive and real-time food safety management, helping to identify risks, anticipate and prevent problems, and trace food sources. This can improve the traceability and management efficiency of the supply chain, ensuring its safety and sustainability.

An intelligent security system is an application system based on artificial intelligence (AI) and information technology, aiming to achieve security monitoring, analysis, and management in specific fields or industries. Intelligent security systems can collect, analyze, and process a large amount of data and information. Technologies such as data mining and machine learning can be utilized to identify potential risks, anomalies, or security threats, and targeted measures can be taken to provide early warning, prevention, and response. Sarker et al. utilized machine and deep learning to extract critical information from raw data, intelligently protecting IoT devices from various network attacks and threats, and establishing an intelligent security system for IoT systems [2]. Gupta et al. proposed graph-based machine learning technology to identify malicious users in intelligent transportation systems, analyze network traffic and detect malicious devices, providing identity authentication for intelligent vehicles in intelligent transportation systems and thereby establishing an intelligent security system for transportation [3]. Yan et al. utilized technologies such as artificial intelligence image recognition, digital maps, IoT in place confirmation, and mobile Internet to establish an integrated infrastructure intelligent security system with comprehensive infrastructure security management as the core [4]. Kumar et al. used load frequency control to simulate virtual inertia in an island power system, and designed a digital frequency relay protection system against network attacks. The researchers implemented intelligent control technology between robust energy storage systems and wind energy systems to establish an island intelligent security system [5]. Walid El-Shafai utilized techniques such as singular value decomposition, discrete wavelet transform, chaotic encryption process, and wavelet fusion algorithm to distribute medical data between two different remote institutions, thus establishing a multi-level intelligent security system for medical color images based on fusion, watermarking, and encryption techniques [6]. Radzi et al. introduced deep learning technology for face recognition and applied the Internet of Things to access control systems (ACS) in universities, using Raspberry Pi as the main controller of the face recognition locking system to establish an intelligent security system [7]. However, due to the numerous links involved in the food supply chain and the large amount of data, the use of the aforementioned security systems cannot process complex data structures, and are inefficient.

In this study, big data analysis technology is used to build an intelligent security system. In the era of Industry 4.0, the significant increase in data through the Internet of Things has led to the rise of the "data-driven" era, and big data analysis has been applied to various industries. The growth of available data is a recognized trend, and valuable information can be derived from data analysis. In this case, big data analysis is an important determinant of competitiveness and innovation in various industries [8]. Tsan-Ming Choi et al. applied big data analysis methods to different thematic areas of modern operations management forecasting, inventory management, revenue management and marketing, transportation management, and risk analysis, and studied the practical application of big data analysis in top brand enterprises [9]. Li Zhu et al. discussed the framework for conducting big data analysis in intelligent transportation systems, and summarized data sources and collection methods, data analysis methods and platforms, and big data analysis application categories. They also introduced the application of big data analysis in intelligent transportation systems, including road traffic accident analysis [10]. Zhang et al. studied the advanced application of big data analysis technology in smart grids. By collecting, storing, and analyzing massive data from power grids, meteorological information systems, geographic information systems (GIS), etc., many benefits were brought to the power system, and customer service and social welfare in the era of big data were improved [11]. Mohammadpoor and Torabi applied big data analysis technology to the oil and gas industry, analyzing seismic and microseismic data to improve reservoir characterization and simulation, analyzing how to shorten drilling time and improve drilling safety, and overcome the current challenges faced by the oil and gas industries [12]. Kumar et al. used big data analysis tools and technologies to process massive heterogeneous data in the healthcare field, exploring the conceptual architecture of medical big data analysis such as genomic databases, electronic health records, text and images, and clinical decision support systems [13]. Wang et al. identified big data analysis technology as the core technology of intelligent manufacturing systems and reviewed related topics such as the concept, model driven, and data driven methods of big data. They discussed the framework, technology, and application of big data analysis in intelligent manufacturing systems, providing new ideas for the implementation of intelligent manufacturing systems [14].

This study uses big data analysis technology to analyze the data generated by various links in the food supply chain and construct an intelligent safety system. This system can help identify potential safety risks, predict and alert safety events, and provide intelligent food safety management and response measures. This current study constructs an intelligent security system for the food supply chain through big data analysis technology. Firstly, sensor data, monitoring data, etc., are collected. HDFS and computing frameworks are used to store, process, and analyze large-scale datasets, and Spark is utilized to process and analyze the data. Visualization tools are used to display security reports and indicators in intelligent security systems, helping users of such systems to better understand and analyze security events and threats, thereby providing decision support. The application of intelligent security systems based on big data analysis in the food supply chain can provide comprehensive food safety management, help identify risks, give warnings of, and prevent, problems, and trace food sources. This can help to improve the traceability and management efficiency of the supply chain, ensuring its safety and sustainability.

Table 1 Part of data information collected by sensors.

Time information	Sensor name	Numerical value	Unit
2022-06-22 12:00	Soil temperature sensor 1	25.3	°C
2022-06-22 12:00	Soil temperature sensor 2	24.9	°C
2022-06-22 12:00	Soil humidity sensor 1	35	%
2022-06-22 12:00	Soil pH sensor 1	4.7	pH
2022-06-22 12:00	Nitrogen (N) sensor 1	160	mg/L
2022-06-22 12:00	Air temperature sensor 1	36.4	°C
2022-06-22 12:00	Rainfall sensor 1	0	mm

Table 2 Information of food storage.

Unique identification code	Time of storage	Delivery time	Type	Quantity	Unit	Batch
UIC20221008759	2022-07-13 06:13	2022-07-15 08:55	wheat	100	kg	B100098
UIC20221357688	2022-07-20 08:34	2022-07-25 13:43	potato	50	kg	B135077
UIC20221443897	2022-07-23 09:05	2022-07-25 10:34	banana	50	kg	B144078
UIC20220987589	2022-07-31 15:33	2022-08-03 18:55	tomato	50	kg	B098045
UIC20221058367	2022-08-03 08:45	2022-08-05 15:50	cucumber	50	kg	B105067

2. DATA COLLECTION

The typical links in a food supply chain include production, storage, transportation, processing, distribution, and sales. When collecting relevant data, different data collection methods need to be adopted based on the characteristics of each link. The current industrial era has integrated the latest digital technologies, and the biggest challenge lies in utilizing these technologies to effectively connect and organize these complex network structures, and identify and meet the needs of food supply chain stake holders [15].

2.1 Collection of Production Data

The collection of production data is done in the production node of the food supply chain. In the food supply chain environment of the platform, data is collected from multiple channels, including the monitoring data from e-commerce platform (sales) and logistics platforms. These data contain diverse information required for the production process. IoT devices are used for the monitoring, control, and collaboration of various objects and systems in the physical world through embedded sensors, actuators, and intelligent devices connected to the Internet. Items in the Internet of Things are analogous to the connection between humans and computers, which can be assigned Internet protocol addresses and transmit data through networks or other artificial means [16].

Regarding agricultural products, data collection is carried out using IoT devices to obtain real-time production parameters and process information. IoT devices can monitor and cooperate with various objects and systems through embedded sensors, actuators, etc. connected to the Internet. In terms of agricultural products, various sensors are used to monitor parameters such as soil, weather, and water quality, and cameras are used to track and detect the growth of agricultural products. Water quality sensors are utilized to monitor indicators such as pH of water quality and water temperature. In this study, production data is collected by means of 20 soil temperature sensors, 20 soil humidity sensors, 10 soil pH sensors, 5 nutrient content sensors, 5 air temperature sensors, 5 air humidity sensors, 5 rainfall sensors, 10 cameras, and 5 pH of water quality sensors arranged in crop planting bases. At specific time points, the values of each sensor can be obtained. The obtained information is shown in Table 1.

2.2 Collection of Storage Data

In the food supply chain, the storage process is of vital importance; hence, temperature is a key factor in preserving food, and humidity is a factor that affects the quality of certain foods. Therefore, it is necessary to monitor the temperature and humidity of the storage area. It is also necessary to record the entry and exit times of food products, calculate their storage time, and monitor their shelf life. The activities carried out to ensure the hygiene of the storage environment must be recorded in terms of procedure and frequency. It is also essential to track the variety, quantity, and batch information of stored food for inventory management and traceability of food sources.

The temperature and humidity data of the storage area are collected through corresponding sensors, and RFID technology can be used to associate food with a unique identification code. RFID tags can be attached to food and unique batch numbers are shown on label information. Information such as time, date, food type, and quantity can be automatically recorded by scanning equipment. The hygiene-related activities require manual input of information. The data collected for the storage process is shown in Table 2.

2.3 Collection of Transportation Data

During food transportation, it is essential to monitor temperature and humidity, and record the position and trajectory of transportation vehicles or cargo containers. Some foods

are sensitive to vibrations during transportation, so it is necessary to add vibration sensors to record the vibration intensity during transportation. It is also important to track the food loading time, transportation start and end time, and arrival time at the destination to ensure that transportation is completed within the specified time. Also, the frequency of cleaning transportation vehicles and containers should be recorded to ensure the hygiene of the transportation process and equipment.

During food transportation, RFID readers can be used to scan labels. Readers and writers can be installed in transportation equipment, reading information on labels through radio waves and recording it. By analyzing the recorded RFID tag's data, the movement trajectory and position information of the food can be obtained, and by comparing the recorded time, position, and operator information, the location of the food can be accurately tracked. The data collected for food during transportation is shown in Table 3.

2.4 Collection of Processing Data

During the processing phase, it is necessary to record data on the input of raw materials, processing steps, time and other production steps, as well as to monitor the temperature and humidity during the processing phase, especially for foods that require temperature and humidity control. This helps to ensure that the food processing is carried out under appropriate temperature and humidity conditions, avoiding quality loss and safety risks. It is vital to record the weight and measurement data of raw materials, finished products, and intermediate products, and to conduct regular sampling and testing of product samples. The recording of sample testing data will ensure that the quality and safety of the product comply with standards and regulatory requirements. It is also necessary to record the frequency of cleaning and disinfecting of equipment, tools, and work areas. Finally, traceability data must be recorded, including the source of raw materials, supplier information, as well as production batch, date, and destination of finished products. Taking flour processing as an example, the time taken for grains to undergo cleaning, grinding, screening, grading, and packing processes is recorded to ensure the quality of food. The collected data information for the processing is shown in Table 4.

2.5 Collection of Distribution Data

The data that needs to be collected for the food distribution process includes the storage status of goods in the distribution center or warehouse, including the type, quantity, inbound and outbound time of the goods. Sales data such as sales orders, sales revenue, sales channels, and customer information need to be recorded. It is necessary to record data such as the delivery time, transportation method, delivery destination, transportation time, route, and cost of the product. Batch data, shelf life information, and related traceability data of products in the distribution process also need to be recorded. The collected data for the distribution process is shown in Table 5.

3. DATA PREPROCESSING

By collecting data from sensors and RFID devices, raw data can be obtained. However, raw data is not appropriate for processing and analysis, so preprocessing technology is applied to the raw data, making it suitable for practical application [17].

3.1 Data Cleaning

Data cleaning is a key step in big data preprocessing. It involves using various methods and techniques to identify and handle incorrect values, inconsistent values, missing values, and outliers in the dataset. Certain methods should be adopted to correct or supplement problematic data to improve the overall reliability of the data [18]. Incorrect values are produced by incorrect data, which may be caused by errors in data input or processing, such as errors between multiple sensors or spelling errors during manual input. Incorrect values are typically addressed through elimination, correction, or estimation. Inconsistent values, arising from input errors or data collection issues, manifest as contradictions within a dataset. These inconsistencies are typically resolved through standardization, data consolidation, and conflict resolution techniques. Missing values, on the other hand, are those that are unrecorded or omitted in observations. Handling missing values often involves deleting records with gaps or interpolating to fill in missing information. Outliers, significantly deviating from normal patterns, can be caused by sensor failures or abnormal data collection conditions. These outliers are commonly treated as missing values or replaced with median or average values to maintain data integrity.

3.2 Data Integration

Data integration involves merging diverse data sources, formats, and storage locations into a unified and comprehensive dataset. Effective data integration not only enhances the efficiency of data access management, but also streamlines scientific data analysis [19]. Methods such as data table connections and column merging facilitate seamless data integration. When integrating data, it is necessary to consider factors such as data structure, data format, and data quality of different data sources to reduce errors and deviations during the data integration process, for example, whether Itemnumber in one dataset is the same entity as Product_id in another dataset. Alternatively, the date and time can be formatted as "Y-m-d H:i" in one system and "Y/m/d H/i" in another system. This requires converting the format to ensure data consistency and comparability.

Table 3 Information for food transportation.

Unique identification code	Time of loading	Time of departure	Time of arrival	Time of discharge	Location	Transport vehicle number
UIC20221008759	2022-07-15 08:55	2022-07-15 10:03	2022-07-16 05:33	2022-07-16 08:38	Datong North Road and Caidi Road intersection	VN10079
UIC20221357688	2022-07-25 13:43	2022-07-25 15:24	2022-07-25 20:45	2022-07-25 23:12	333 Beiying Road, Qingpu District	VN10079
UIC20221443897	2022-07-25 10:34	2022-07-25 13:13	2022-07-25 19:48	2022-07-25 22:06	29 Huangshan West Road	VN10378
UIC20220987589	2022-08-03 18:55	2022-08-03 20:32	2022-08-04 07:36	2022-08-04 10:22	736 Yangzijiang Middle Road	VN10279

Table 4 Information for food processing.

Unique identification code	Weight (kg)	Cleaning time	Grinding time	Screening time	Grading time	Packing time
UIC20221008759	100	2022-07-16 10:00	2022-07-16 10:30	2022-07-16 12:55	2022-07-16 14:24	2022-07-16 16:30
UIC20221007896	50	2022-07-16 10:00	2022-07-16 10:30	2022-07-16 12:55	2022-07-16 14:24	2022-07-16 16:30
UIC20221006879	30	2022-07-16 10:00	2022-07-16 10:30	2022-07-16 12:55	2022-07-16 14:24	2022-07-16 16:30
UIC20221008972	30	2022-07-17 10:00	2022-07-17 10:28	2022-07-17 13:02	2022-07-17 14:20	2022-07-17 16:33

Table 5 Product distribution process data.

Item number	Type	Weight(kg)	Time of storage	Order number	Batch
IN20359	flour	160	2022-07-20 12:34	ON202207201342	B10001
IN20488	oil	100	2022-07-20 14:45	ON202207202458	B20013
IN20156	flour	100	2022-07-21 09:22	ON202207214581	B10023
IN20568	rice	100	2022-07-22 10:48	ON202207222768	B30245

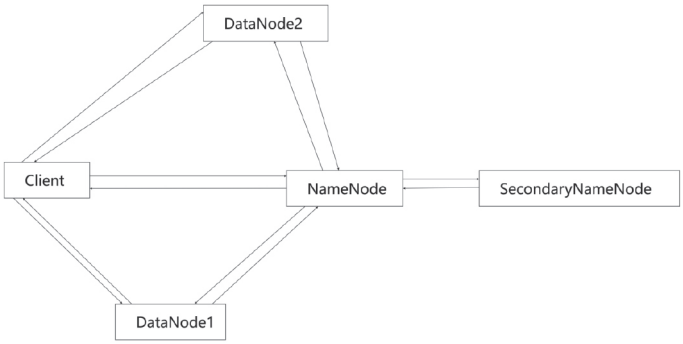

Figure 1 Working structure of HDFS components.

3.3 Data Conversion

Data conversion refers to the transformation of raw data, which is the most complex and difficult problem in data preprocessing [20], that is, how to transform data to change its distribution, scale its range, so as to make it more in line with the requirements of the analysis model. Standardization methods can be used to scale data to a specific range, commonly including interval scaling and unit length scaling, to ensure that different variables have the same weight. Data conversion can improve the quality and adaptability of data, thereby improving the accuracy and reliability of subsequent analysis and modeling.

4. BIG DATA STORAGE AND MINING

After preprocessing the data for each link in the food supply chain, the data can be stored. Hadoop Distributed File System (HDFS) was chosen as the tool for storing food supply chain data on the platform. HDFS is a distributed file system suitable for processing large-scale data, which has high reliability and scalability, and can meet the needs of food supply chain data processing.

Today's rapid development of internet technology has seen the emergence of Google and Yahoo, which are typical types of data-intensive applications that use big data-oriented infrastructure to provide scalable services. The quality of service of HDFS determines the reliability and performance of these applications [21]. HDFS generally divides file systems into two parts: data and metadata. Compared to traditional distributed file systems, HDFS has two important advantages. The first is significant fault tolerance, as it can save copies of data in multiple data clients, allowing for recovery of data when different errors occur in other data clients [22]. The second benefit is that it can scale horizontally to thousands of servers, providing very high storage capacity and storing petabyte level data. HDFS has been widely used in recent years due to its advantages such as easy deployment, low operating costs, high reliability, scalability, and big data processing [23].

4.1 HDFS Cluster Construction

HDFS is a Master/Slave structure [24], which mainly consists of DataNode, NameNode, SecondaryNameNode, and Client. DataNode is the working node of HDFS, responsible for the actual storage and processing of data. Each DataNode manages locally stored data blocks and regularly sends heartbeat signals to NameNode, reporting its own storage information and status. DataNode is also responsible for copying data blocks and processing client requests. NameNode is one of the key components of HDFS. It is responsible for managing and controlling the namespace of the file system, metadata information of files, and location information of data blocks throughout the cluster. NameNode is a single point of failure, so its reliability is very important. For large-scale clusters, backup Secondary NameNodes can be used to assist NameNodes in metadata backup and recovery operations. Secondary NameNode is not a backup node for NameNode, but a secondary node. It is responsible for regularly taking snapshots of the file system from NameNode and helping with metadata consolidation and organization operations, reducing the burden on NameNode. SecondaryNameNode can improve the reliability and performance of NameNode. Client is a user or application that interacts with HDFS. It provides a set of APIs (Application Programming Interface) and command-line tools, enabling users to read, write, delete files, and manage file systems. The working structure of the four components of HDFS is shown in Figure 1.

In Figure 1, there is only one NameNode and Secondary-NameNode, while there are multiple DataNodes. In this study, 90 Linux servers were prepared in the production,

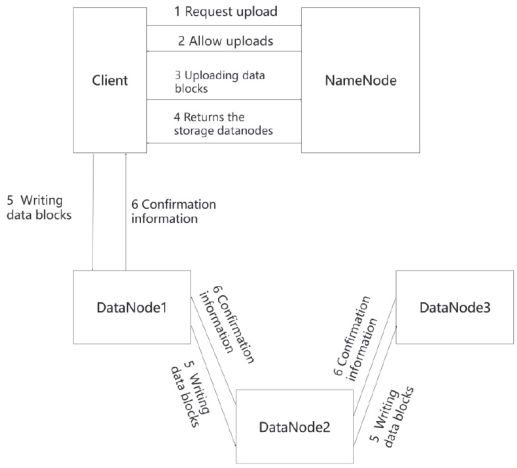

Figure 2 Data uploading process.

storage, transportation, processing, and distribution stages of the food supply chain, and all servers in the cluster need to be interconnected. So, to configure the network environment well, the IP address and subnet mask are included, and the host name of each server is added to the/etc/hosts file. Each server can access and communicate with others through host names. The HDFS cluster involves multiple nodes. If people want to start the HDFS software processes on each node in the cluster in bulk, they need to use a batch startup script. Here, SSH (Secure Shell) is used for password-free login, which sends remote commands to other nodes to start corresponding processes through SSH. The vast majority of software components in the entire Hadoop technology ecosystem are based on Java. Therefore, all nodes in the cluster must install JDK (Java Development Kit) and configure corresponding environment variables. Finally, the configuration time synchronization service can be installed to ensure the normal operation of the HDFS cluster.

4.2 HDFS Upload Data

Firstly, the collected data needs to be converted into a format supported by HDFS through a script. In this paper, the data is converted into JSON (JavaScript Object Notation) format, and then the data is uploaded using the HDFS API. The uploading process is shown in Figure 2.

The client first communicates with the NameNode, requesting to create a file in HDFS, and NameNode indicates whether it can be uploaded. The client then requests which DataNodes to upload the data block to, and the NameNode returns the DataNode. Then the client finds the corresponding DataNode and requests an upload of the data. The DataNode continues to call the other two DataNodes, and the three DataNodes respond to the client level by level, ultimately completing the upload. This design approach achieves concurrent access, reduces addressing overhead, and greatly improves data access speed [25].

Among the most advanced parallel computing platforms, Spark is a fast, universal, and in-memory iterative computing framework. It is used for large-scale data processing, ensuring high fault tolerance and scalability by introducing Resilient Distributed Datasets (RDD) [26]. Firstly, the Spark client submits tasks to the Spark cluster, and then the Spark task reads data from the data source HDFS during execution. It can load data into memory, convert it into RDD, and then call some higher-order functions for RDD to process the data, ultimately writing the calculation result data back to HDFS. The intelligent security system selects a Spark computing framework with strong applicability, which can support programming in multiple types of languages. In addition, it can also achieve interactive batch processing and computational analysis of big data [27]. Spark implements the reading and calculation of local partitioned data, thereby reducing the interactive transmission of cluster node data and further improving the system's efficiency in analyzing big data [28].

4.3 Visualization

This article exports the data in Spark to a format, such as CSV (Comma Separated Values). It then saves the exported file in an accessible location so that the visualization tool can read the data file, and uses the file reading function provided by the visualization tool to read the exported data file. It can choose a style for output based on the data format to achieve data visualization. In big data processing, data visualization can better understand the structure and characteristics of datasets, provide a more intuitive way to display the distribution and patterns of data, and thus better understand data and make decisions.

4.4 Big Data Mining

Big data mining technology involves identifying and aggregating relevant data within a specific range, synchronously detecting and classifying them [29]. Big data mining technology is a powerful tool for analyzing massive amounts of data in the food supply chain. It is useful for identifying potential risks and issuing early warning, to achieve the traceability of product and raw material sources. Spark is a leading platform that supports a variety of data mining algorithms, including k-means, which helps detect unusual

patterns or behaviors in the supply chain. In addition, the graphical database effectively tracks the origin and movement of food, providing important insights.

4.4.1 K-Means Algorithm

The k-means algorithm is a clustering algorithm that is used to divide a set of data into k different clusters, resulting in the highest similarity of data points within the cluster. The similarity of data points between clusters is the lowest. It requires specifying the initial number of clusters and initial clustering centers in advance based on the distance between samples, and then dividing the sample set into clusters based on the similarity between objects and clustering centers. It continuously updates the position of the cluster center and reduces the Sum of Squared Error (SSE) of the cluster. When the SSE no longer changes, the clustering ends and the final result is obtained.

Firstly, this article randomly selects k initial cluster centers $C_i(i \leq 1 \leq k)$ from the dataset, calculates the Euclidean distance between the remaining data objects and the cluster center C_i, identifies the cluster center C_i closest to the processing target data object, and assigns the data object to the cluster corresponding to the cluster center C_i. Then, the average value of the data objects in each cluster can be calculated as the new clustering center for the next iteration, until the clustering center no longer changes or the maximum number of iterations is reached. The Euclidean distance calculation formula between data objects and clustering centers in space is:

$$d(X, C_i) = \sqrt{\sum_{j=1}^{m}(X_j - C_{ij})^2} \tag{1}$$

Among them, X is the data object, C_i is the i-th clustering center, m is the dimension of the data object, X_j, C_{ij} are the jth attribute values of X and C_i. The SSE calculation formula for the sum of squared errors of the entire dataset:

$$\text{SSE} = \sum_{i=1}^{k}\sum_{X \in C_i} |d(X, C_i)|^2 \tag{2}$$

Among them, k is the number of clusters.

To use the k-means algorithm in Spark, the first step is to use the processing tool RDD provided by Spark to load the raw data into Spark. The feature processing tools and algorithms provided by Spark ML are used to process features to ensure that appropriate feature vectors are obtained. For example, in food production processes, transportation methods, etc., before clustering, it is necessary to normalize the features to ensure that different features have the same scale. The StandardScaler tool provided by Spark can be used to normalize features, and then the k-means in Spark ML is used for clustering. After clustering is completed, the cluster center is extracted from the k-means model, which represents the feature vector of each cluster and can be used as the result of feature extraction. For new data samples, the trained k-means model can be used to convert them into metric values equal to the Euclidean distance from the clustering center, as the feature extraction results of the new samples. Finally, the data samples can be divided into different categories based on the clustering results, such as storage conditions, hygiene conditions, etc. It can also classify the safety risk categories of the food supply chain, such as high risk, medium risk, and low risk.

Through cluster analysis, food can be classified according to different characteristics and attributes, helping food supply chain managers specify corresponding quality standards and management measures for different categories of food. It can classify suppliers according to their different management levels, production technologies, and other characteristics, help supply chain managers identify high-quality suppliers, and establish a risk management system. Risk assessment can be conducted based on quality inspection reports, test results, and historical data of different categories of products or suppliers, and corresponding risk management measures can be taken, such as product recalls, supplier penalties, and adjusting food safety strategies. Monitoring points, monitoring periods, monitoring objects, monitoring indicators, etc., can be classified according to different links in the food supply chain. This can optimize monitoring plans and management strategies to ensure food safety.

4.4.2 Graph Database

The tracing of food via Spark is achieved through a graph database. There is a correlation between food production and consumption behavior in the food supply chain. When analyzing data, in order to better utilize association relationships, graphs are often used as data structures, and databases that use graph structures to store data are called graph databases. Traditional relational databases display data in a table structure, making it easy to query and manage data. The graph database focuses more on the connections between nodes and surrounding nodes, which is a mesh structure suitable for traceability analysis.

The graph in the database consists of three parts:

1) Point represents an entity object in the graph, represented as a node in the graph. For example, people and food can be abstracted as a node in the graph.

2) Edges are the relationships between nodes in a graph, such as the purchasing behavior of food.

3) Attributes are used to describe the attributes of nodes or edges in a graph, such as number and name.

Spark can be used to build a graph database by first loading the graph dataset into Spark. Depending on the data format, Spark's DataFrame or RDD provided API can be used to load data. Based on the loaded node and edge data, Spark's Graph Computing Framework (GraphX) or Graph Processing Framework (GraphFrames) can be used to construct graphs. These frameworks provide APIs and functions for creating graph objects to represent and manipulate graph data in Spark. For a more intuitive understanding, in this study, the information was transformed into images based on the expression of the graph database. The graph structure is shown in Figure 3.

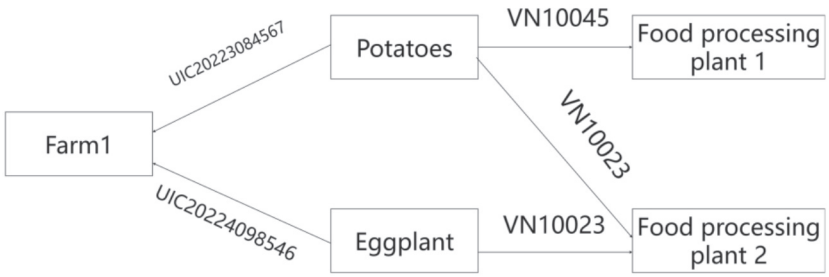

Figure 3 Data structure.

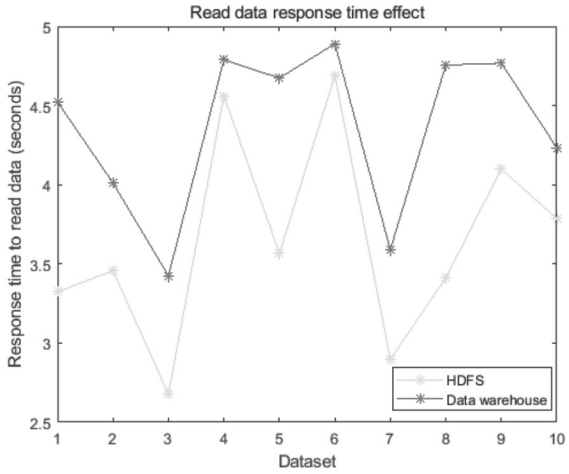

Figure 4 Read data response time.

After establishing the graph node information, Gremlin can be used for querying. Gremlin is an open source streaming query language with flexible query implementation. GraphX is Spark's graph computing framework that provides a set of APIs allowing users to program and query distributed graphs using Gremlin language. By using this API, Gremlin queries can be used to retrieve data from the graph database, perform graph traversal and computation operations. In Figure 3, the relationship between the farm, agricultural products, and processing plant entities is clearly presented, which is more convenient for correlation analysis. For example, it is easy to find the processing plant to which potatoes are sent, and other foods that are also sent to that processing plant. The graph database makes it very convenient to query the information in surrounding nodes, facilitating traceability.

5. EXPERIMENTAL RESULTS

After data collection and preprocessing, a total of 10365 files were obtained on the HDFS cluster, resulting in 2469800 pieces of data. These data were divided into 8 training datasets and 2 test datasets for experiments. The experimental results are presented below.

5.1 Evaluation of Experimental Results

5.1.1 HDFS Response Time Effectiveness Evaluation

HDFS provides readily-available and flexible storage services for clusters. As a storage entity, when the disk usage in the cluster is high, the read and write load on DataNodes is relatively high. This can easily cause the data synchronization task to fail due to read/write timeout, so it is necessary to evaluate its response time for writing and reading files. Response time and throughput are also two important indicators used for measuring the performance of big data analysis and traditional data processing methods. Figure 4 shows the response time results for the reading of data. Figure 5 shows the response time results for the writing of data.

From Figure 4 and Figure 5, it can be seen that there is not much difference in response time between HDFS and the data warehouse when reading data. HDFS has a faster response time, with an average of 3.647 seconds. When writing data, HDFS is much faster than traditional data warehouses, with an average response time of 6.139 seconds.

5.1.2 Evaluation of HDFS Throughput Effect

HDFS achieves parallel processing of massive data and has high error tolerance characteristics. At the same time, it has the advantages of high throughput, which is the amount of work completed per unit of time. Figure 6 shows the throughput of HDFS when reading and writing data.

As shown in Figure 6, HDFS has a higher throughput compared to data warehouses, with an average of 896MB/s, indicating that HDFS can process massive amounts of data very efficiently.

5.1.3 SSE Curve in K-Means

This minimizes SSE by finding an optimal segmentation method. The more novel the SSE, the better the model

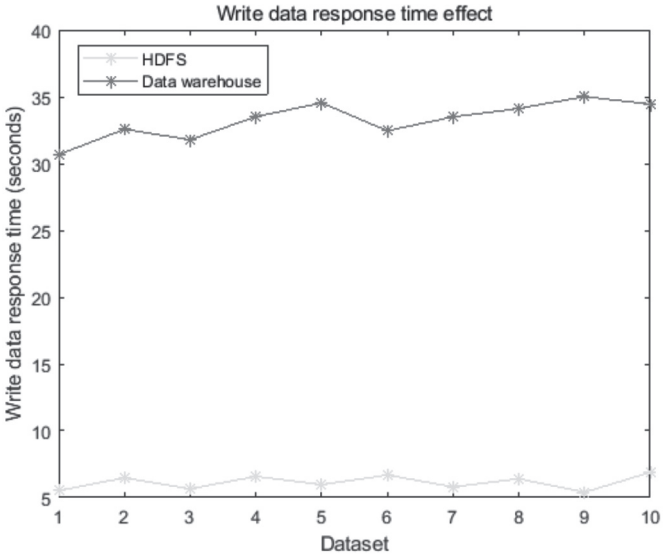

Figure 5 Response time for writing data.

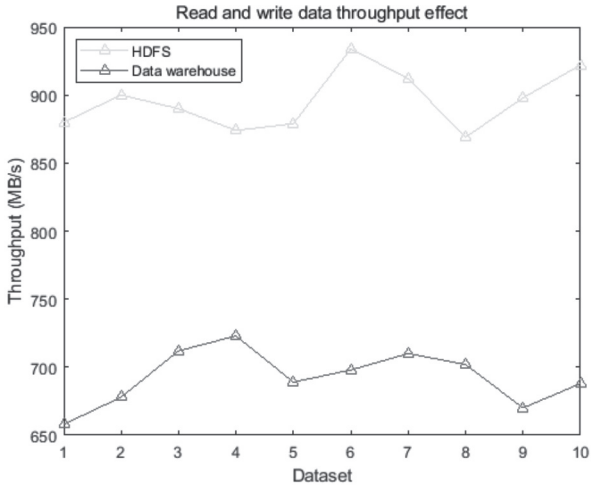

Figure 6 HDFS read and write data throughput.

selection and fitting, and the more accurate the data prediction. Therefore, the SSE curve can be used to evaluate the prediction performance of the system. Figure 7 shows the SSE values gradually decreasing as the number of clusters increases.

5.2 Conclusion of Experimental Results

The experimental results show that using HDFS to read and write data has a faster response time and higher throughput. It can achieve more efficient processing of massive data. At the same time, the clustering analysis algorithm used in Spark also has relatively accurate prediction results, which helps to achieve a comprehensive and intelligent security system.

6. DESIGN AND IMPLEMENTATION OF INTELLIGENT SECURITY SYSTEM

Security is a comprehensive concept that involves multiple aspects, and a security system is a highly technical and comprehensive project. The core of an intelligent security system should be the information center and monitoring center [30]. The intelligent safety system of the food supply chain can monitor, identify, prevent, and respond to risks and safety issues throughout the entire food supply chain.

6.1 Temperature Monitoring

Fresh agricultural products such as vegetables and fruits should be kept at a relatively low temperature during transportation, usually between 5°C and 10°C, to extend their shelf life. The intelligent safety systems in the food supply chain platform is applied to the temperature monitoring process. By monitoring in real time the temperature of products during transportation, abnormal temperature conditions can be detected in a timely manner, such as temperature increases caused by refrigeration system failures or incorrect operation of the transportation vehicle. When the temperature does not meet the standard for a long time, the system will automatically issue a warning and trigger emergency

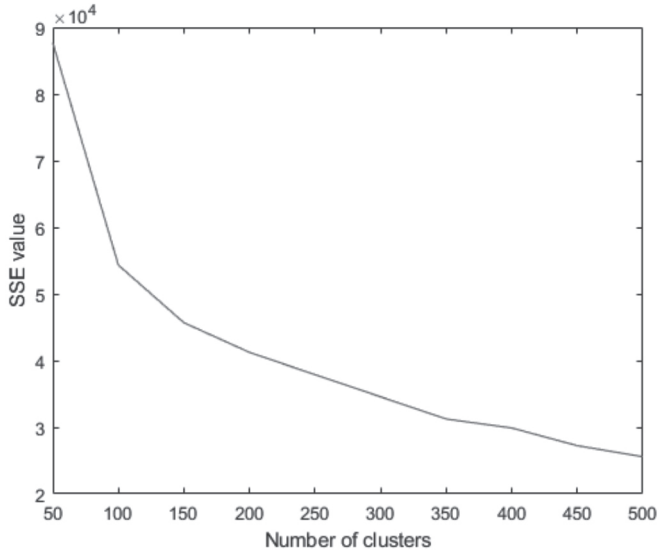

Figure 7 SSE curve.

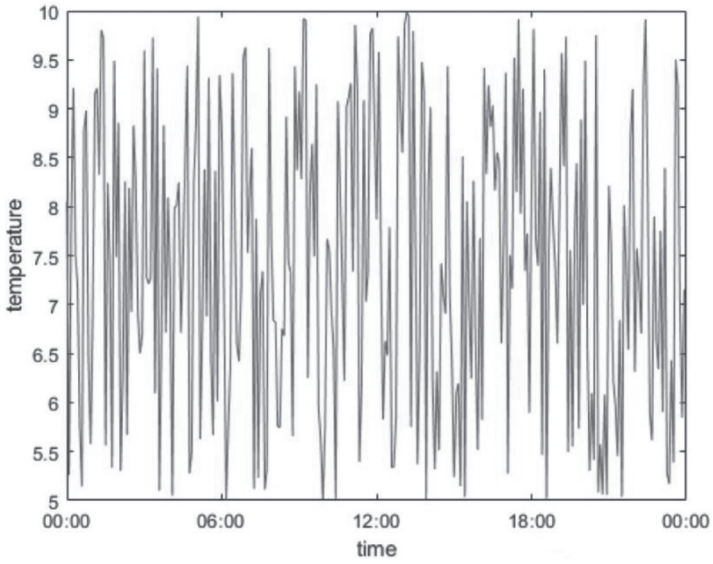

Figure 8 Temperature inside the transport vehicle.

response measures. The system can automatically send alert notifications to relevant responsible persons, record abnormal situations, and trigger the quality inspection process. By detecting the temperature during transportation, the freshness of food can be determined. When the temperature is above what is required for a certain length of time, it is necessary to trace back to this batch of food for quality testing. The temperature monitoring effect is shown in Figure 8.

6.2 Hygiene and Cleanliness Monitoring

The hygiene and cleanliness of food storage environments are crucial, requiring regular cleaning and disinfection to eliminate bacteria and pollutants. When the frequency of hygiene cleaning decreases or the cleaning process is not detailed, the hygiene of the storage area must be closely inspected to reduce the risk of food pollution, bacterial growth, and food spoilage, and ensure the safety and quality of food. As an example, this study took a flour processing factory where it is necessary to clean the floor and equipment of the processing plant every day at 6:00 am, 12:00 noon, and 18:00 pm. The floor is cleaned and disinfected as is all of the processing equipment. After completion, the operator records and signs off on the hygiene and cleanliness activities as shown in Table 6.

6.3 Transportation Time Tracking

The tracking of food transportation time is a crucial step in ensuring the safety and quality of food being stored and transported. The barcode of a food product helps to track its status at every stage of the supply chain. When the transportation time is too long, relevant personnel can be notified of food abnormalities using SMS or other methods. The tracking effect of transportation time is shown in Table 7.

Table 6 Hygiene and cleanliness table.

Name	Cleaning method	Operator	Operating time
Ground	Clean-disinfect	N0001	06:00
Equipment	Wipe-disinfect	N0001	
Ground	Clean-disinfect	N0002	12:00
Equipment	Wipe-disinfect	N0002	
Ground	Clean-disinfect	N0001	18:00
Equipment	Wipe-disinfect	N0001	

Table 7 Transportation schedule.

Unique identification code	Departure time	Departure point	Time of arrival	Place of arrival	Transport vehicle	duration
UIC20221003478	2022-03-25 8:00	Wuxi City	2022-03-25 20:00	Shanghai Municipality	VN20045	1 day
UIC20221006785	2022-03-25 8:00	Wuxi City	2022-03-25 20:00	Shanghai Municipality	VN20045	1 day
UIC20222006745	2022-04-01 10:32	Wuhan City	2022-04-02 8:00	Wuxi City	VN10067	2 days
UIC20222008912	2022-04-01 14:12	Changsha City	2022-04-02 11:34	Wuxi City	VN20034	2 days
UIC20223008791	2022-04-05 13:56	Wuxi City	2022-04-06 08:00	Hangzhou City	VN20054	2 days
UIC20223009867	2022-04-07 08:23	Nanjing City	2022-04-07 15:33	Suzhou City	VN10056	1 day
UIC20223001923	2022-04-09 18:23	Nanjing City	2022-04-10 09:38	Jinan City	VN20134	2 days

Figure 9 Traceability of food sources.

6.4 Food Source Traceability

Firstly, food labels and packaging can provide information on the source and production of the food, including production date, production location, manufacturer information, batch number, etc. Consumers can know the source and production background of food by reading the information on labels and packaging. With big data analysis, food traceability information can be obtained by scanning or querying traceability codes, such as planting location, breeding location, and destination city. Figure 9 shows the traceability information of food sources.

As shown in Figure 9, the origin of the food is Chenzhou City, passing through Hengyang City, Xiangtan City, Yiyang City, and finally reaching Yueyang City.

7. CONCLUSIONS

The application of intelligent security systems based on big data analysis of the food supply chain has yielded promising results. Through intelligent security systems, data on food production, transportation, processing, storage, and distribution can be monitored and traced in real-time. This study analyzed the occurrence of patterns in food safety issues. The intelligent security system of the food supply chain still has some shortcomings as the system needs to collect and integrate data from multiple links so as to obtain complete supply chain information. Due to the interests of individual parties in a supply chain, there may be misleading data due to false positives or concealments, which will affect the accuracy of the system. In the future, a better and

more accurate intelligent safety system should be designed to address the existing problems so as to improve the safety of the food supply chain and its efficiency. This would give the end consumers more reliable food safety information and promotesustainable development of the food supply chain.

FUNDING

This work was supported by the National Social Science Fund of China (20BGL129).

REFERENCES

1. J. Astill, RA. Dara, M. Campbell, et al., Transparency in food supply chains: A review of enabling technology solutions. Trends in Food Science & Technology, 91 (2019), 240–1247.
2. IH. Sarker, AI. Khan, YB. Abushark, et al., Internet of things (iot) security intelligence: a comprehensive overview, machine learning solutions and research directions. Mobile Networks and Applications, 28(1) (2023), 296–312.
3. BB. Gupta, A. Gaurav, EC. Marin, et al., Novel graph-based machine learning technique to secure smart vehicles in intelligent transportation systems. IEEE Transactions on Intelligent Transportation Systems, 24(8) (2022), 8483–8491.
4. HP. Yan, W. Zhang, XH. Shi, et al., Infrastructure security system based on intelligent image recognition. Automation Expo, 38(07) (2021), 70–74.
5. N. Kirn Kumar, V. Indra Gandhi,Logesh Ravi, V. Vijayakumar, V. Subramaniyaswamy. Improving security for wind energy systems in smart grid applications using digital protection technique[J]. Sustainable Cities and Society, 60(2020): 102265.
6. W. El-Shafai, EED. Hemdan, Robust and efficient multi-level security framework for color medical images in telehealthcare services. Journal of Ambient Intelligence and Humanized Computing (2021), 1–16.
7. S. Radzi, M. Alif, YN. Athirah, et al., IoT based facial recognition door access control home security system using raspberry pi. International Journal of Power Electronics and Drive Systems, 11(1) (2020), 417.
8. K. Vassakis, E. Petrakis, I. Kopanakis. Big data analytics: Applications, prospects and challenges. Mobile big data: A roadmap from models to technologies (2018), 3–20.
9. TM. Choi, SW. Wallace, Y. Wang, Big data analytics in operations management. Production and Operations Management, 27(10) (2018), 1868–1883.
10. L. Zhu, FR. Yu, Y. Wang, et al., Big data analytics in intelligent transportation systems: A survey. IEEE Transactions on Intelligent Transportation Systems, 20(1) (2018), 383–398.
11. Y. Zhang, T. Huang, Ef. Bompard. Big data analytics in smart grids: a review. Energy Informatics, 1(1) (2018), 1–124.
12. M. Mohammadpoor, F. Torabi, Big Data analytics in oil and gas industry: An emerging trend. Petroleum, 6(4) (2020), 321–328.
13. S. Kumar, M. Singh, Big data analytics for healthcare industry: impact, applications, and tools. Big Data Mining and Analytics, 2(1) (2018), 48–57.
14. JL. Wang, CQ. Xu, J. Zhang, et al., Big data analytics for intelligent manufacturing systems: A review. Journal of Manufacturing Systems, 62 (2022), 738–752.
15. M. Lezoche, JE. Hernandez, MMEA. Diaz, et al., Agri-food 4.0: A survey of the supply chains and technologies for the future agriculture. Computers in Industry, 117 (2020), 103187.
16. AA. Laghari, K. Wu, RA. Laghari, et al., A review and state of art of Internet of Things (IoT). Archives of Computational Methods in Engineering (2021), 1–19.
17. J. Luengo, GG. Diego, RG. Sergio, et al., Big Data Preprocessing. Cham: Springer (2020), 1–14.
18. GD. Huang, ZH. Long, ZP. Zhu, et al., Monitoring data analysis of water supply pipe network based on support vector machine. Water Supply and Drainage, (48) (2022), 124.
19. SB. Cheng. Discussion on a technology management data integration platform based on big data. Technological Style, (35) (2022), 63–65.
20. L. Li, G. Dan. A brief discussion on big data integration. Fujian Computer, 35(01) (2019), 162.
21. T. Ma, F. Tian, B. Dong. Ordinal optimization-based performance model estimation method for HDFS. IEEE Access 8, (2019), 889–899.
22. HK. Omar, AK. Jumaa, Big data analysis using apache spark MLlib and Hadoop HDFS with Scala and Java. Kurdistan Journal of Applied Research, 4(1) (2019), 7–114.
23. GR. Chen, XH. Yuan, Real-time detection algorithm for abnormal data based on HDFS open source architecture. Computer Simulation, 38(8) (2021), 445–449.
24. QR. Wei, Design of hospital information system based on HDFS architecture. Information and Computers (Theoretical Edition), 35(04) (2023), 133–135.
25. FX. Shi, Y. Gao, Application analysis of Hadoop big data technology. Modern Electronic Technology, 44(19) (2021), 153–157.
26. RX. Guo, Yi. Zhao, Q. Zou, et al., Bioinformatics applications on apache spark. GigaScience, 7(8) (2018), 98.
27. B. Ning. Design and implementation of Spark-based big data analysis system. Information Recording Materials, 24(09) (2023), 202–204.
28. YH. Chen, Design and development of multi-dimensional data analysis system for cultural tourism in the big data environment. Electronic Testing, (4) (2021), 62–64.
29. XX. Ge, Application of big data mining technology in network security. Digital Technology and Applications, 41(07) (2023), 225–227.
30. Y. Wang, SY. Wu, Design and implementation of intelligent community security system. Anhui Construction, (06) (2005), 13.
31. P. Zhao, A. Jensen, T. Johnsen, "Blockchain Ecosystem Meet Supply Chain Ecosystem and an Application to Dairy Product Provenance", Engineering Intelligent Systems, vol. 32 no. 1, pp. 19–23, 2024.
32. F. Cui, M. Ni, T. Zhou, H. Wang, "Key Technology Research on End-Side Arithmetic Network Based on Resource Virtualization for Multi-Terminal Systems", Engineering Intelligent Systems, vol. 31 no. 5, pp. 379–387, 2023.

Xin Zhang was born in Dazhou, Sichuan P.R. China, in 1973. He received the Master from Wuhan University of Technology, P.R. China. Now, he studies in College of Business, Jiaxing University. His research interest include SCM and logistics big data analysis.
E-mail: Zcxin9@163.com