ISSN: 2710-0723

Agricultural Product Data Analysis System based on Big Data Platform

Tao Jin, Enze Wu, Taizhi Lv

College of Information Engineering, Jiangsu Maritime Institute, Nanjing, Jiangsu 211170, China

Abstract

With the continuous development and popularization of information technology, the agricultural sector has gradually entered the digital age. The traditional agricultural data processing methods are relatively inefficient and cannot meet the growing data demands and complex agricultural production environments. This article constructs an agricultural product data analysis system, which automatically obtains data such as agricultural product prices and market information by crawling and parsing data from agricultural product related websites. By utilizing big data platforms for processing and analysis, statistical analysis, mining, and modeling of agricultural product market data are conducted, providing users with a comprehensive display and prediction of national agricultural product price trends, market size, hot selling products, and other data. This ensures the reliability and efficiency of the data, and utilizes visualization technology to provide users with a user-friendly interface and convenient data query functions.

Keywords

Agricultural Product Data; Spark; Hive; Echarts.

1. Introduction

With the continuous development and popularization of information technology, the agricultural sector has gradually entered the digital age. The collection, processing, and analysis of agricultural data have become pivotal to the advancement of modern agriculture [1]. In 2023, the total grain output reached 6954.1 billion kilograms, an increase of 8.88 billion kilograms or 1.3% over the previous year, and remained stable at over 0.65 trillion kilograms for nine consecutive years. Among them, the output of corn was 288.84 billion kilograms, an increase of 4.2%; The soybean output was 20.84 billion kilograms, an increase of 2.8%. The relevant data is increasing[2]. However, the traditional agricultural data processing methods are often inefficient and cannot meet the increasing data demand and complex agricultural production environment[3][4]. Therefore, it is crucial to build an efficient agricultural product data analysis system. The establishment of the Agricultural Data Analysis System can help agricultural practitioners better understand market demands and trends, and improve the efficiency and quality of agricultural production and sales [5] [6]. For government departments and decision makers, ADAS also provides important data support, which helps to formulate more scientific and effective agricultural policies and development strategies. Therefore, research and development of Agricultural Data Analysis System has important practical significance and far-reaching social impact [7].

At present, some agricultural product data analysis systems have emerged at home and abroad [8], but there are still some problems and deficiencies. Some systems have a single function, providing only simple data display and query functions, lacking in-depth data analysis and mining capabilities; Other systems lack flexibility and scalability, and cannot adapt to the constantly changing agricultural data environment. Therefore, it is necessary to study and

ISSN: 2710-0723

improve the existing agricultural product data analysis system to enhance its functionality and performance.This paper designs and implements a powerful, flexible and scalable agricultural product data analysis system.

2. System Design

2.1. Functional Modules

As shown in Figure 1, the Agricultural Product Data Analysis System primarily comprises functional modules such as Data Management, Data Visualization, Price Prediction, and Price Inquiry. The implementation of these functions necessitates the utilization of various technical means, including Batch Processing Map APIs, positioning technology, data storage technology, linear regression, and more.

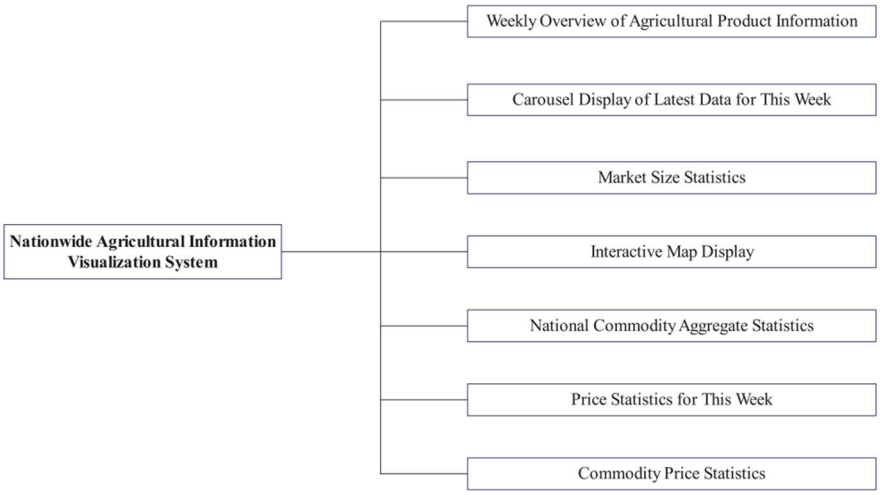

Figure 1. Functional Module Diagram

2.2. System Architecture

This system consists of a Data Acquisition Layer, Data Processing Layer, Data Storage Layer, and Data Visualization Layer. The Data Acquisition Layer utilizes crawler technology based on Selenium and urllib to obtain agricultural product price data and price trend data from designated websites. The Selenium crawler simulates browser operations to capture agricultural product price data and stores it in CSV files. By simulating real user actions, the Selenium crawler can scrape dynamically loaded content, ensuring data integrity and real-time accuracy. This method is particularly suitable for handling content that requires user login or interaction, ensuring timely data updates. The urllib crawler sends HTTP requests to obtain agricultural product price trend data and stores it in JSON format. By directly requesting APIs or static pages, the urllib crawler efficiently retrieves large-scale data and ensures standardized data formats. This approach is simple and efficient, suitable for obtaining well-structured data sources.

The Data Processing Layer is responsible for data processing using pandas, reading agricultural product price data from CSV files, and formatting it. It also includes parsing and formatting JSON data, reading data from CSV and JSON files using Spark SQL for further processing and statistics. SQL statements are used to aggregate, group, and sort data to obtain the required statistical information. Pandas can read, clean, transform, and merge CSV and JSON data, ensuring data consistency and integrity. Through the data frame, the system can easily perform data filtering, selection, and transformation operations for subsequent analysis. Spark SQL is used for processing and analyzing large-scale data, improving data processing efficiency through

ISSN: 2710-0723

distributed computing. Spark SQL can handle massive amounts of data and perform complex aggregations, groupings, and sorting operations using SQL statements to generate detailed statistical reports.

The Data Storage Layer is primarily responsible for storing processed data and information. The acquired and processed CSV files, JSON files, and analyzed data are stored in the Hive database, supporting subsequent queries and analysis. The Hive database is deployed in virtual machines with VMware's high availability status, including managers to ensure system availability and stability. The Data Visualization Layer is responsible for presenting processed data to users in a visual manner. It is implemented using a front-end and back-end separation architecture, with the front-end providing data interfaces and services based on the Spring Boot framework, and the back-end implementing page displays based on HTML, CSS, and JavaScript. The Data Visualization Layer utilizes the Mapbox API to visualize map data, displaying geographical location information and spatial distribution of agricultural product markets. Through the collaborative work of these layers, the system efficiently completes the acquisition, processing, storage, and visualization of agricultural product price data, providing price trend prediction functionality and comprehensive data services and decision support for users. The system architecture is designed reasonably, with independent yet closely cooperating layers, ensuring the system's efficiency, reliability, and scalability. The system not only meets current data processing and analysis needs but also has good expansion capabilities to cope with future increases in data volume and complexity. The system architecture is shown in Figure 2.

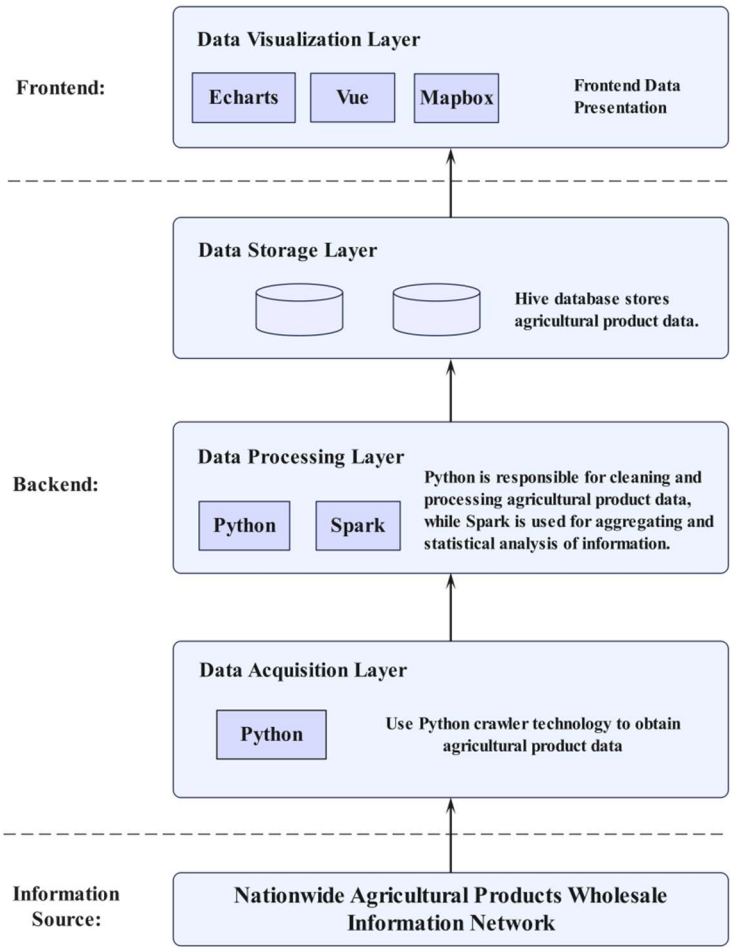

Figure 2. System Architecture Diagram

ISSN: 2710-0723

3. System Design and Implementation

3.1. Background Data Processing

The backend data processing section mainly comprises three parts: data acquisition, processing, and storage. The data acquisition part utilizes web crawling technology to obtain agricultural product price data and price trend data from specified websites by simulating browser operations and sending HTTP requests. The crawled data is stored locally in CSV and JSON formats. The data processing part uses the pandas library and Spark SQL engine. The pandas library is employed to read data from files, format and merge it for subsequent statistics and analysis. Meanwhile, the Spark SQL engine processes and generates statistics on data from CSV and JSON files, producing corresponding data results. The data storage part is responsible for saving the processed data in persistent storage to ensure data reliability and efficiency. Storage is managed through a Hive database, catering to various data management and query requirements across different scenarios.

3.2. Implementation of Frontend Visualization

The data visualization section employs a front-end and back-end separated architecture. The back-end provides data interfaces and services based on the Spring Boot framework, while the front-end implements page presentation using HTML, CSS, and JavaScript. Additionally, the Mapbox map API is utilized to visualize map data, displaying the geographical location information and spatial distribution of agricultural product markets. The overall system effect is shown in Figure 3.

Figure 3. Overall System Effect Diagram

In the center of the visualization webpage, a MapBox map is used as the base map, with geographical locations of provincial capitals displayed on it. Each location marker is bound with a click event, and clicking on a marker retrieves agricultural product data for that city and displays it in a pop-up window. This allows users to view and understand data for various cities through the map. Additionally, interactive functionality is provided, enabling users to explore the data more deeply.

It also includes a data display section, which retrieves data from the Hive database, converts it into a list of DataItem objects, and returns it as a JSON response to the client. As shown in Figure 4, it alternately displays detailed information about each product. Additionally, it compiles statistics on the market size of each province in China and displays the top eight provinces with

ISSN: 2710-0723

the largest market sizes. When the mouse is hovered over a corresponding province, it will display the market size of that province, as shown in Figure 5.

Figure 4. Carousel for Commodity Data

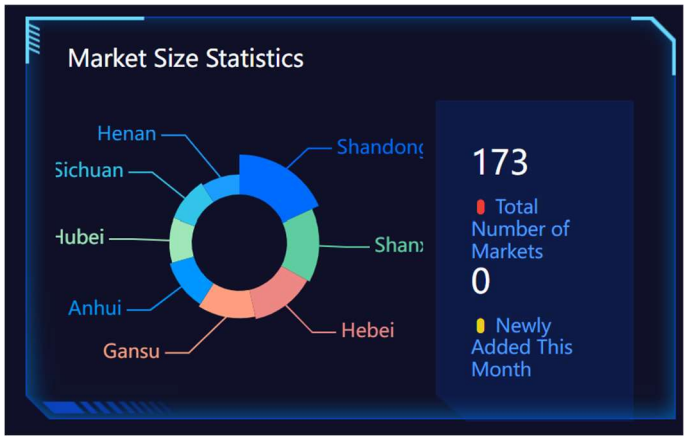

Figure 5. Market Size Statistics

4. Conclusion

The Agricultural Product Data Analysis System is a technological support that leverages modern information technology and data analysis methods to provide intelligent assistance for agricultural production and agricultural product management. The system aims to collect, process, and analyze data related to agricultural products, offering effective decision-making support and management tools for farmers, government departments, and agricultural enterprises. The Agricultural Product Data Analysis System realizes functions such as data prediction, querying, and map visualization. Users can manage agricultural product information, conduct data querying and analysis, and visually present agricultural product-related data through the map interface.

With the continuous development of technology, the Agricultural Product Data Analysis System will become more intelligent, automated, and practical. Here are some potential future development directions:

ISSN: 2710-0723

The prediction module currently considers too few types of data, only price data. When predicting future data, it should incorporate other factors as references, such as seasonal factors, weather data, soil data, policy data, supply and demand data, etc.

Introduce artificial intelligence technology or more advanced machine learning to improve the accuracy of prediction results.

Some visual charts in the current system have significant caching issues, leading to slower responses and delayed display of chart information. In the future, the backend caching mechanism will be optimized, and data will no longer be directly retrieved but instead accessed through Redis caching technology to reduce data extraction time.

Acknowledgments

This work was financially supported by the funding of the Philosophy and Social Science Research Project of the Jiangsu Higher Education Institutions of China (2022SJYB0804), the Fundamental Computer Education and Teaching Research Project of the Association of Fundamental Computing Education of Chinese Universities (2024-AFCEC-416), and the Excellent Teaching Team for QingLan Project of the Jiangsu Higher Education Institutions of China (Big Data Technology Teaching Team with Shipping Characteristic).

References

[1] Meiling, Hu . "Big Data Mining and Analysis of Agricultural Products Based on e-Commerce Platform." Wireless Communications & Mobile Computing 2022(2022).

[2] Osinga, S. , et al. "Big data in agriculture: Between opportunity and solution." Agricultural Systems (2022).

[3] Lutsii, Oleksandr , O. Helevei , and V. Zhuk . "Formation of Components of the Marketing Information System for Agricultural Products Using Big Data Methods." Accounting & Finance / Oblìk ì Fìnansi 101.3(2023).

[4] Guo, Wei , and K. Yao . "Supply Chain Governance of Agricultural Products under Big Data Platform Based on Blockchain Technology." Scientific programming 2022.Pt.1(2022):4456150.1-4456150.16.

[5] Tian, Tian , Y. Zhang , and Y. Mei . "Intelligent analysis of precision marketing of green agricultural products based on big data and GIS." Earth Science Informatics (2022).

[6] Su, Zhifang , Q. Li , and J. Xie . "Based on data envelopment analysis to evaluate agricultural product supply chain performance of agricultural science and technology parks in China." custos e agronegocio on line 15.1(2019):314-327.

[7] Jiang, Leilei , and W. Sun . "Analysis of Agricultural Product Marketing Channels Based on Diversity under the Background of Big Data." Journal of Physics Conference Series 1574(2020):012119.

[8] Shi-Wei, X. U. . "Agricultural Big Data and Monitoring and Early Warning of Agricultural Products." Journal of Agricultural Science and Technology (2014).