Design and Implementation of Educational Data Analysis System Based on Hive and Spark

Yongling Liu [a], Xuejun You [b], Taizhi Lv [c]

School of Information Engineering, Jiangsu Maritime Institute, Nanjing 211170, China;

[a]594697472@qq.com, [b]99026743@qq.com, [c]404050569@qq.com

Abstract

With the continuous increase in the scale of students in universities and colleges, a massive amount of educational data has been accumulated in daily teaching activities. Faced with this explosive growth of data, traditional educational systems primarily focus on business management, lacking analysis of the educational data. This vast amount of data contains potential values about students' learning situations, the status of the teaching team, evaluations of teaching quality, educational research, and innovations, among many others. To fully tap into the value embedded in the educational administration data and display it in an intuitive graphical format, this paper designs and implements an educational data analysis system based on Hive and Spark technologies. It effectively assists universities in better evaluating teaching, optimizing teaching content, and compensating for the shortcomings of traditional educational administration systems.

Keywords

Educational Data, big data technology, Spark, Hive, data visualization.

1. Introduction

With the development of the era, big data technology has been widely applied in various fields. The empowerment of the educational administration system through big data has also attracted the attention of many scholars [1-2]. Cui researched how to improve the functionality and application effects of the educational system through big data technology, enhancing the efficiency and quality of educational administration. Hu etc. studied methods to present educational data and information using data visualization techniques. The research aims to help educators, decision-makers, and other relevant individuals better understand and analyze educational data through charts, graphics, and interactive interfaces to support decision-making and improve educational management.

To fully tap into the value embedded in educational big data and improve teaching quality, this paper designs and implements an educational big data analysis and visualization system based on Hive and Spark. The system presents and analyzes administrative data visually, helping education administrators, teachers, and students better understand and utilize administrative information. The system employs the open-source Hadoop distributed storage as its underlying architecture and sets up the Hive data cleaning platform. By invoking the Sqoop data migration tool, the source data from the Oracle database is migrated to Hive for data cleaning. Using the Spark data processing engine connected to Hive for data analysis, the statistical analysis results are stored in the MySQL database. Finally, the analysis results are visualized through front-end and back-end separation technology. Utilizing data mining and analysis techniques, the system explores and discovers hidden patterns, trends, and associative rules in the educational data, providing support for educational decision-making and teaching improvements.

2. Requirement Analysis and Design

2.1. Requirement Analysis

As shown in Figure 1, the entire system is divided into the index, student data statistics, teacher data statistics, and study data statistics. The index page analysis includes the total number of students, the top ten failure rates of student courses, teaching lectures, and so on. The student data statistics is analyzed from four perspectives: sex ratio, age segmentation, college student distribution, and annual student statistics. The teacher data statistics is examined from four aspects: teacher teams, teacher positions, the distribution of teacher numbers across colleges, and age segmentation. The study data statistics is analyzed based on student academic performance, the number of scholarship recipients, the number of certificate recipients, and the segmentation of grade point averages (GPA). These analysis results can help schools identify some teaching issues, improve teacher training and student support, enhance teaching quality and student engagement, optimize resource allocation and utilization, avoid resource wastage and shortages, and increase the efficiency of educational resource usage.

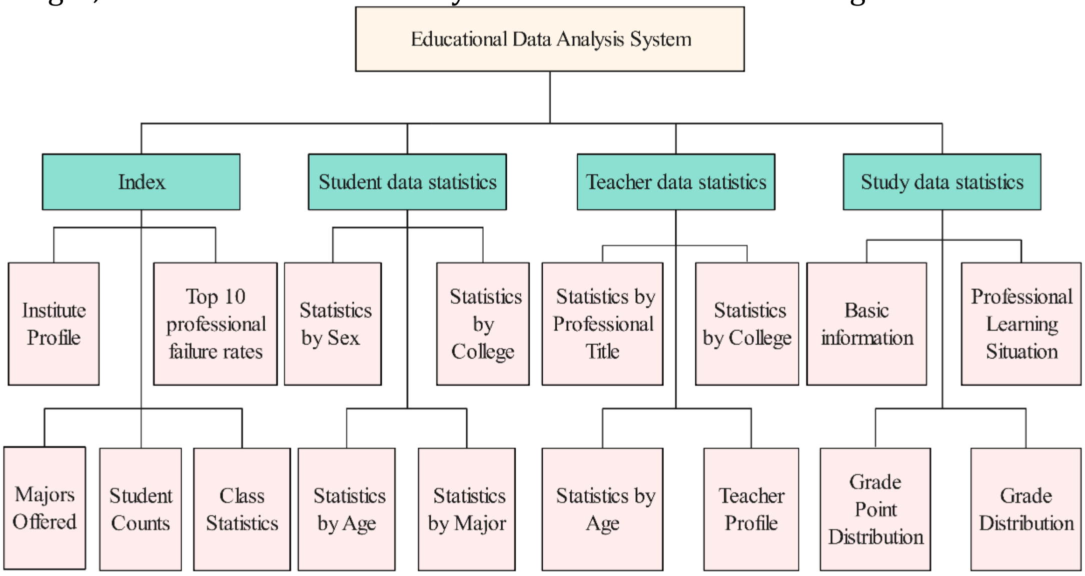

Figure 1. The diagram of the functional module

2.2. Functional Design

As illustrated in Figure 2, the entire system architecture comprises the data source layer, the data warehouse layer, and the data application layer.

The data source layer stores educational administration data and is implemented using the Oracle database. This layer primarily encompasses tables such as score tables, course tables, basic student information tables, teaching schedules, and basic teacher information tables. Metadata is extracted into the Hive data warehouse using ETL tools. The data warehouse layer is realized by Hive, responsible for offline data analysis and statistics. The results of the Hive data analysis are exported to the MySQL database using Sqoop. The data application layer is implemented with the MySQL database and a visualization program. The MySQL database stores the statistical results, and the visualization program is made possible through front-end and back-end separation technology.

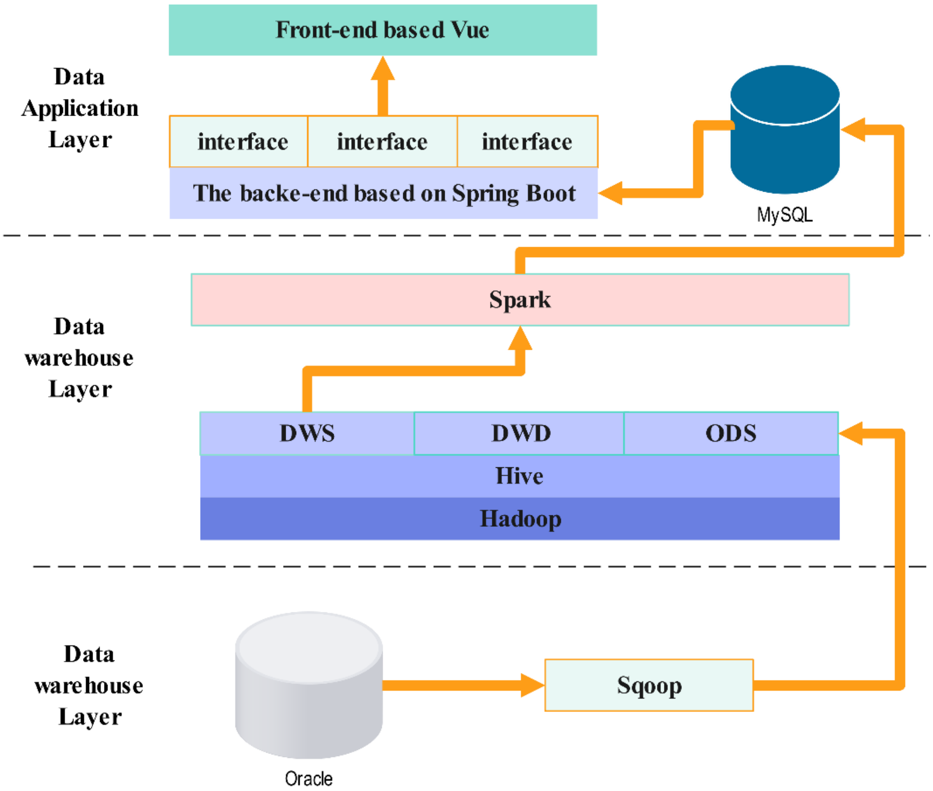

Figure 2. The architecture of the system

3. System implementation

3.1. Data collection

Data collection is accomplished using the open-source tool Sqoop. Sqoop is a tool designed for transferring data between Hadoop and relational databases. It fundamentally relies on the MapReduce framework and collects data in bulk from relational databases through JDBC connections [5]. The extraction of administrative data is done in two steps. Initially, a full extraction is done, loading all the data into the data warehouse. Subsequent operations employ incremental extraction, updating the data warehouse with any new data from the operational database, and manipulating the data warehouse by monitoring the database logs.

3.2. Hierarchical Data Storage Based on Hive

Hive is a data warehousing tool built upon Hadoop, capable of extracting, transforming, storing, and querying massive datasets. Hive offers an SQL querying interface, allowing SQL statements to be transformed into MapReduce tasks for execution [6]. This system utilizes Hive to construct a three-tier architecture for educational data analysis: the Data Operation Layer, Data Detail Layer, and Data Service Layer. Through layered testing, issues can be promptly identified and addressed, minimizing the risk of latent issues propagating upwards. This approach facilitates concentrating testing focus on the specific layer and its adjacent layers. By doing so, test cases are drafted and executed more rigorously, ensuring precise testing of big data.

In the Data Operation Layer, educational data from the source, Oracle, is stored intact. This data serves as the foundation for subsequent processing. When importing from Oracle to Hive, all char and varchar types are converted to string types, number types to double types, while date and TIMESTAMP types remain unchanged. The Data Detail Layer is responsible for cleaning the data from the ODS layer and dimensionally reducing and modeling the business data tables to reduce inefficiencies caused by multi-table joins in subsequent queries. The Data Service Layer, from various thematic perspectives, aggregates and summarizes data from the Data Detail Layer, producing statistical results.

3.3. Data Cleaning Based on Hive and Data Analysis with Spark

After collection, Hive is used to cleanse, filter, and write the educational data into the Data Operation Layer. Data tables undergo quality checks to identify data issues and anomalies, including missing values, outliers, duplicated records, and inconsistent data formats. Utilizing Hive's query statements and built-in functions allows for swift preprocessing of the data.

Spark is a fast and general-purpose big data processing and analysis engine. It excels in handling large-scale datasets, offering efficient data processing capabilities [7]. Leveraging Spark's distributed computing power, massive educational data can be efficiently processed and analyzed. This facilitates statistical analysis of the educational data, ensuring accurate and reliable results.

3.4. Data Visualization Based on Front-end and Back-end Separation

To better showcase the data analysis results, this study employs a front-end and back-end separation technique to visualize the analytical findings. The back-end, based on Spring Boot, retrieves the necessary fields from the database using SQL queries, returning the data in JSON format. Vue then uses AXIOS to send asynchronous requests, acquiring various types of data returned by the back-end. This data is rendered onto the web pages as tables, charts, or layers on maps.

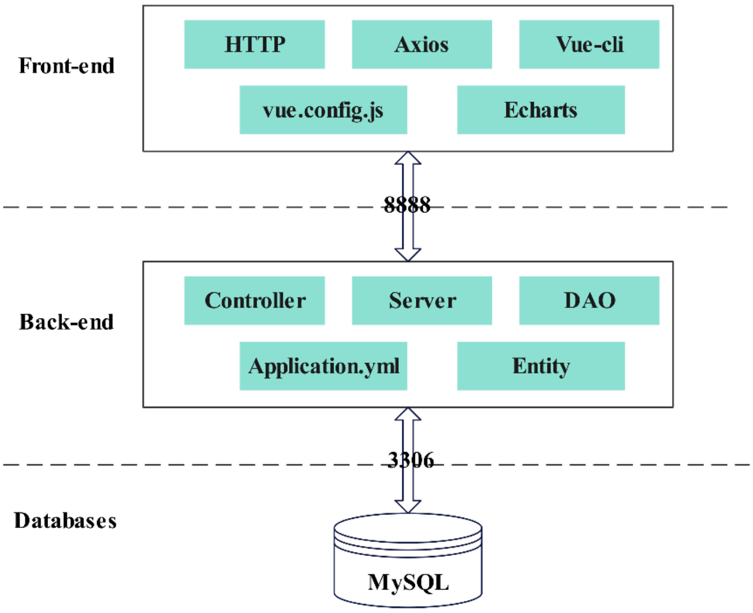

Figure 3. The architecture of the data application layer

The visualization module's backend, based on Spring Boot, is implemented in three layers. The first layer, the Controller, provides data for the front-end visualization module, implemented using annotations. The second layer, the Service, focuses on operations addressing specific issues. It retrieves data from various databases and handles the logic, such as caching, computations, and business rules. The third layer, the DAO layer, mainly operates on data within the MySQL database, dealing with data CRUD (create, read, update, delete) operations.

To simplify the development of the visualization layer, the front-end is implemented using Vue and is compiled and packaged using Webpack. Vue is the premier open-source development kit for the MVVM (Model-View-ViewModel) pattern [8]. It facilitates two-way data binding, allowing for the incremental construction of user interfaces. Chart displays are implemented using Echarts. ECharts is a JavaScript library for visualizing data, offering intuitive, vibrant, interactive, and highly customizable visual charts. Data for the charts is fetched using Axios asynchronous requests. Once retrieved, the data is bound to the corresponding Echarts chart object, which then renders the appropriate chart on the associated container.

4. Conclusion

With the advancement of the times, the development of society, and the refinement of information technology, more and more industries are using information technology to aid their growth. At present, university administrative data is vast and intricate. Without timely analysis of data fluctuations, many courses aren't arranged logically, leading to student learning issues. To address these challenges, an educational big data visualization module is designed. This system utilizes the Spark+Hive big data platform to analyze data. The back-end employs the Spring Boot framework to supply data. The front-end fetches data via AJAX, binds data with Vue, and achieves visualization with Echarts, offering a clear display of relevant educational data.

Acknowledgements

This work was financially supported by the funding the Philosophy and Social Science Research Project of the Jiangsu Higher Education Institutions of China (2020SJA0730), the 13th Five-Year Plan Project of Education Science in Jiangsu Province(B-a/2018/03/22), the Fundamental Computer Education Research Project (2021-AFCEC-285) of the Association of Fundamental Computer Education Research Project of China, 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] Duan, Yanqing, John S. Edwards, and Yogesh K. Dwivedi. "Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda." International journal of information management 48 (2019): 63-71.

[2] Baig, Maria Ijaz, Liyana Shuib, and Elaheh Yadegaridehkordi. "Big data in education: a state of the art, limitations, and future research directions." International Journal of Educational Technology in Higher Education 17.1 (2020): 1-23.

[3] Cui, Libiao. "Construction of big data technology training environment for vocational education based on edge computing technology." Wireless Communications and Mobile Computing 2022 (2022): 1-9.

[4] Hu, Xuefeng, and Su Yanhua. "Vocational Education Curriculum Evaluation Model Based on Big Data." Solid State Technology 63.5 (2020): 10070-10078.

[5] Zou, Yujuan, et al. "Research on ship data analysis based on Spark platform." 2023 5th International Conference on Communications, Information System and Computer Engineering (CISCE). IEEE, 2023: 249-252.

[6] Lv, Taizhi, Jun Zhang, and Chenyong He. "Research on posts analysis based on data process automation." 2021 2nd International Symposium on Computer Engineering and Intelligent Communications (ISCEIC). IEEE, 2021: 156-159.

[7] Lv, Taizhi, Juan Zhang, and Yong Chen. "The research on data acquisition and analysis platform for lathe machine based on stream computing." Journal of Physics: Conference Series. Vol. 1650. No. 3. IOP Publishing, 2020: 032060.

[8] Song, Junhui, Min Zhang, and Hua Xie. "Design and implementation of a vue. js-based college teaching system." International Journal of Emerging Technologies in Learning (Online) 14.13 (2019): 59.