A distributed framework of student data early warning system

Chunhong Qian [1, *], Zongwu Xie [2]

[1] School of Accounting, Tongji Zhejiang College, Jiaxing, China

[2] Polytechnic institute, Zhejiang University, Hangzhou, China

* Corresponding Author Email: qianchunhong@tjzj.edu.cn

Abstract. With the deepening of research on big data, various intelligent technologies have been widely used in the digital construction of campus. But the current campus students early warning system can not achieve high accuracy results. It is necessary to collect big data of students and perform online analysis with advanced techniques to timely predict problems of learning performance and activity abnormality. In this paper, the campus big data early warning system is designed based on the distributed framework. It integrates the data collection, storage, mining and analysis, visualization and message notification. The system consists of six layers of data application structure and one security control platform, and uses multiple linear regression algorithm and logistic regression algorithm to predict the abnormal behavior of student, Hadoop and Spark are used to build a distributed framework to monitor student abnormalities.

Keywords: Distributed Framework; Hadoop; Spark; early warning.

1. Introduction

With the completion of "13th Five-Year Plan of Education Informatization" and "Action Plan of Teaching Informatization 2.0", the Smart Campus which integrates artificial intelligence, big data, cloud computing and other technologies has been widely applied in the digital construction of campus. However, it is still difficult to build a comprehensive campus big data warehouse due to the wide scope of business, many application systems, and no unified format. As a result, big data technology does not play its role in school teaching management. Especially as the most important subject of campus education, students' campus activity and academic performance are the core of school teaching management, their timely warning of abnormal problems is the key to the healthy growth of students' body and mind, but abnormal problems of students are not found in time due to the lack of original data and low timeliness of collection. Therefore, it is the most urgent and worthy of study to construct the big data warehouse of school teaching management. Hadoop [1] and Spark [2] are currently popular computing frameworks that focus on data storage and parallel computing, so based on distributed framework, the abnormal warning system of student data is constructed.

2. Distributed Framework

Data intelligent collection technology, data analysis technology and data visualization technology provides the possibility to solve the problems of data collection and intelligent analysis of campus student activity management. However, the application of these three technologies alone cannot solve the problems of real-time management and dynamic early warning. Therefore, it is necessary to integrate the data collection, storage, mining and analysis, visualization and message notification to build a big data early warning system. Data collection and storage can be solved by the Hadoop ecosystem, and data statistics and mining analysis can be solved by the Spark ecological framework.

Hadoop is a distributed system infrastructure developed by the Apache Foundation, which can make full use of cluster computing to develop distributed programs. Hadoop has evolved into a big data ecosystem with distributed collaboration service (ZooKeeper), real-time log collection (Flume), ETL (Sqoop) and data warehouse processing tool (Hive).

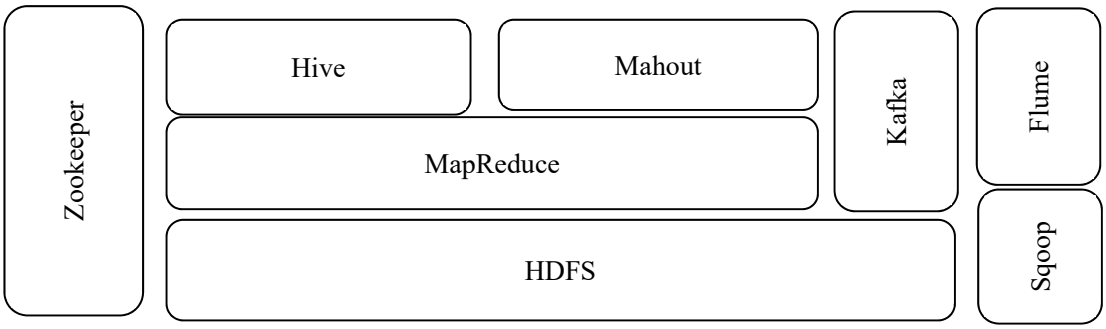


Figure 1. Hadoop ecosystem

Spark is a fast and universal computing engine designed for large-scale data processing. Its computation is based on memory and the read and write speed is faster, so it is more suitable for iterative MapReduce algorithms in data mining and machine learning. Spark has evolved into a big data computer mining and processing ecosystem that includes Spark SQL, Spark Streaming, MLlib (Machine Learning Library), and Graphx (Graph).

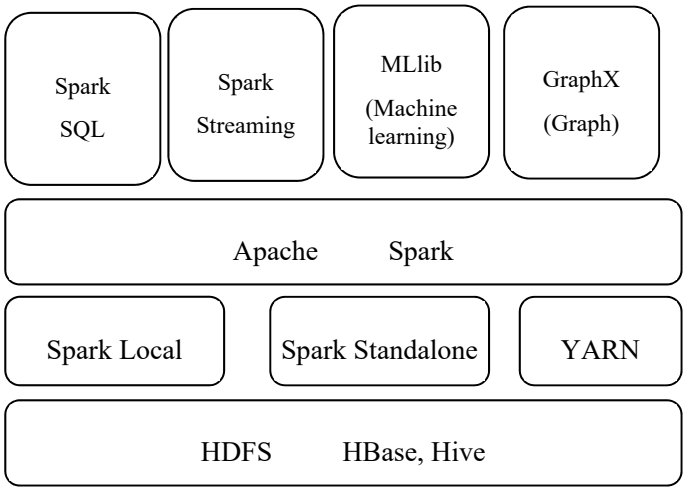


Figure 2. Spark ecosystem

3. Systematic Architecture Design

The early warning system of campus student activities includes the collection, processing and storage of students' campus activities and academic performance data to the big data warehouse. Data mining is applied to analyze early warning and the results are visualized in the human-computer interface. The concrete thoughts are: Firstly, through face recognition, human body and behavior recognition, OCR character recognition and other visual intelligent recognition technology, the system can complete the digitized and the incomplete digitized activity data of students for the full quantitative collection. Secondly, in the data mining technology, RSME evaluation score prediction model and F1-score index evaluation activity anomaly warning model are used to select the best model for dynamic anomaly warning [3]. Finally, whether the abnormal warning system can play the maximum role, it needs to apply data visualization technology to present it. According to the applicable scenarios of "Application Program Visualization development Tool" and "Web Application Data Visualization Development Tool", ECharts can be selected to complete the front-end visualization display.

In terms of technical architecture, it applies the Hadoop big data collection and storage framework to meet the requirements of full quantitative primitive data collection and distributed storage, applies

Spark big data computing framework to meet the timeliness requirements of big data mining and analysis, and applies MVC three-tier architecture to complete Web application development, and Stable Build to complete Wechat small program application development.

3.1 System framework Design

The early warning system is made of the "6 + 1" basic framework design which includes six layers of data application structure and one security management and control platform. The six layers of data application structure: data acquisition layer, data processing layer, data storage layer, data mining analysis layer, data visualization layer, front-end application layer. One security management and control platform: data security, data scheduling, permission control, log monitoring, data lineage, metadata management, data quality management, and data disaster recovery processing. As shown below:

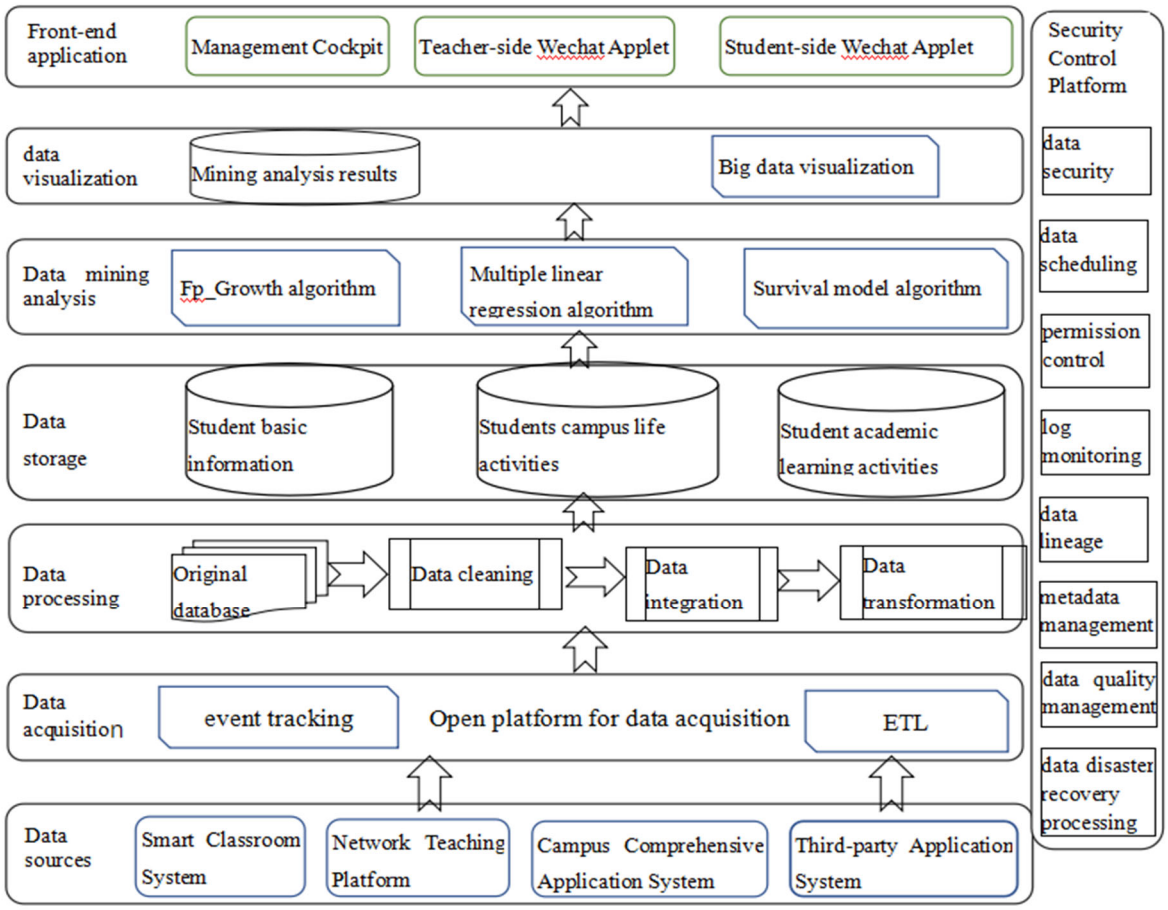


Figure 3. System framework design drawing

The data acquisition layer collects the original data from multiple application modules which includes smart classroom system, network teaching platform and campus comprehensive application system and the third-party application system. The data processing layer includes the cleaning, integration, quantification and transformation of raw data. The Data storage layer is divided into three data warehouses: student basic information, students campus life activities and student academic learning activities. The data mining analysis layer includes calculation of anomaly warning results of big data statistical analysis and data mining algorithm. The data visualization layer presents the warning results using visual expression coding. The front-end application layer is divided into management cockpit, teacher-side Wechat applet and student-side Wechat applet.

3.2 Technical Architecture Design

The technical architecture is divided into three layers: original data acquisition source, big data acquisition, storage, mining and analysis layer, front-end application development. The original data collection source stores the data in system logs or MySQL database through back-end WebAPI and visual intelligence awareness. The big data acquisition, storage, mining and analysis layer adopts Flume+ Kafka and Sqoop to collect the original data in each application system to Hadoop distributed system in real time. Spark SQL performs statistical analysis of big data, Spark Spark MLlib mines data and stores the results to the MySQL database. The application program adopts WebAPI technology and RabbitMQ technology to push the warning message to the front end, the small program development framework is presented with Stable Build and ECharts, and the Web side uses VS2019 development framework and Vue+ECharts to present.

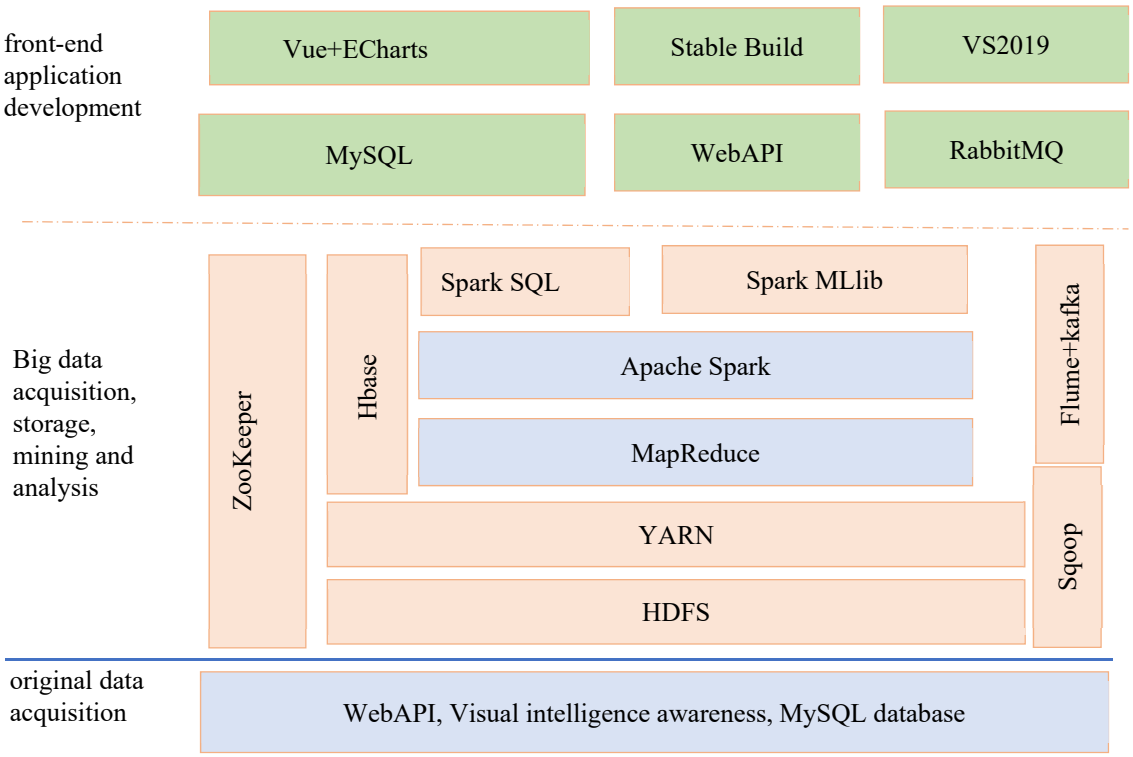


Figure 4. System technical architecture

3.3 Data Warehouse Construction

According to the classical hierarchical construction principle of data warehouse, the Data Warehouse of student activities and academic performance is divided into Operation Data Store layer (ODS), Data Warehouse Detail layer (DWD), Data Warehouse Service layer (DWS) and Application Data Store layer (ADS). ODS is used to store all the data originally collected, DWD stores the cleaned and summarized data of ODS layer, DWS stores the data wide table of student activities and academic performance, ADS stores the data of mining statistical analysis results.

4. Conclusions

Students' activity safety and academic performance are the core of campus management. This paper designs a dynamic early warning system based on the big data of campus activities. The "companion" + "ETL" data acquisition method is proposed and implemented, which can effectively collect students' activity data of various network applications on campus. A big data warehouse is constructed after performing data cleaning, integration, and transformation processing. Then we use

multiple linear regression and logistic regression algorithm to predict abnormal behavior of students. This provides an effective implementation path for improving the management level of school students.

With the establishment of national education standardized data warehouse, improvement of technology, and the emergence of open data sets of education, more and more education big data mining and analysis applications will emerge to serve school management [4]. As an important component of school teaching management, abnormal warning of campus student activities is a direction worthy of long-term research.

References

[1] Erin Gilheany. Processing time of TFIDF and Naive Bayes on Spark2.0, Hadoop 2.6 and Hadoop 2.7: Which Tool Is More Ecient? [D]. School of Computing National College of Ireland, 2016: 1-18.

[2] Iskender Ulgen Ogul, CANER Ozcan.Fast text classification with Naive Bayes method on Apache Spark [C] Signal Processing and Communications Applications Conference (SIU). 2017: 164-169.

[3] Vaculíková,Jitka.Measuring self-regulated learning and online learning events to predict student academic performance[J]. Studia paedagogica, 2018, 23(4): 91-118.

[4] Pistilli, M.D, ARNOLE, K.E. Purdue signals: mining real-time academic data to enhancestudent success [J]. About Campus: Enriching the student learning experience, 2010(3).