Design and Research of Online Education Data Platform Based on Hadoop

Huan Zhou[1](✉), Yu Xu[2], Chunling Ding[1], and Cuiyun Wang[1]

[1] College of Big Data and Artificial Intelligence, Anhui Xinhua University, Hefei, China
zhouhuan0813@ustc.edu.cn

[2] School of Computer Science and Technology, University of Science and Technology of China, Hefei, China
xuyun@ustc.edu.cn

Abstract. Online education is a convenient way to combine content dissemination and rapid learning with electronic products and Internet. Online education advantages include high content production efficiency, no time and space constraints, and study whenever or wherever you want. During COVID-19 pandemic, millions of students have switched to online learning, and online education platform has meet new development opportunities. This thesis specifically analyzes online education platform based on Hadoop through big data. Online education data platform includes four panels, such as access and consultation panel, user registration panel, user intention panel, and student attendance panel. Panel analysis results show that the number of registered users converted from intended customers, attrition rate and attendance of existing users. It helps us to accurately increase user number of online education platform. This provides reliable data support and important reference value for sustainable development of online education.

Keywords: Online Education · Sqoop · Relational Database · Hadoop

1 Introduction

Online education is a kind of learning behavior based on Internet. Its main products are selling video courses and setting up online study rooms to manage learning methods of users so as to improve their learning efficiency. Nowadays, many courses in colleges and universities have adopted the mode of online education MOOC, which not only saves teacher resources, but also enables public people to solve problems of uneven distribution in some educational resources [1, 2].

By the end of 2020, user scale of online education reached 342 million, and usage rate of online education was 34.6 percent. Online education is developing rapidly, but there are still various problems, such as easy bottlenecks, low sustainable development. Therefore, how to increase user number and improve user education quality is a huge problem facing online education. Therefore, based on Hadoop, the construction of online education data platform and visual data analysis provide important reference value for the development and programing of this industry [3, 4].

X. Yuan et al. (Eds.): ICEKIM 2023, AHCS 13, pp. 1802–1813, 2023.
https://doi.org/10.2991/978-94-6463-172-2_200

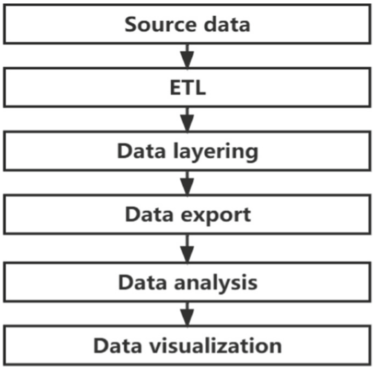


Fig. 1. Data platform structure [self-drawing]

2 Design and Layering of Data Platform

2.1 Data Platform Structure

Data platform can truly reflect online education user consultation, registration, intention, and student attendance. It is divided into four panel areas, which are access and consultation panel, user registration panel, user intention panel, and student attendance panel.

Figure 1 shows data platform design structure. Three core processes of data platform are layering, exporting and visualization. The working process of data platform is: (1) Import source data; (2) ETL collects source data; (3) Data is stored and layered; (4) Data analysis and visualization; (5) Summarize conclusions through visualization results [5].

2.2 Data Platform Design

The purpose of data platform is to retain the information of intended users who visit the platform and convert it into registered users, which can improve the utilization rate of online education platform. At the same time, it can also highlight the advantages of online education platform, improving learning quality of students and ensuring stable retention rate [6].

Data layering is to standardize the processing and transformation of all data, determine data processing and data application steps. This paper introduces data layering to solve data problems. Figure 2 shows that data layering can clarify data structure, reduce repeated development, facilitate maintenance, unify data interfaces, and simplify complex problems.

(1) ODS: Original data layer. It belongs to first layer of data platform; (2) DWD: Detail data layer, which is result data after ETL/ELT processing of original data; (3) DWM: Intermediate data layer. Classify and process according to time dimension (year, month, day, hour), attribute dimension (regional information, campus, discipline), data source (0 is offline, 1 is online), and customer attribute (0 is old customer, 1 is new customer). Data tables are associated with each other using primary key “id”; (4) DWS/DM:

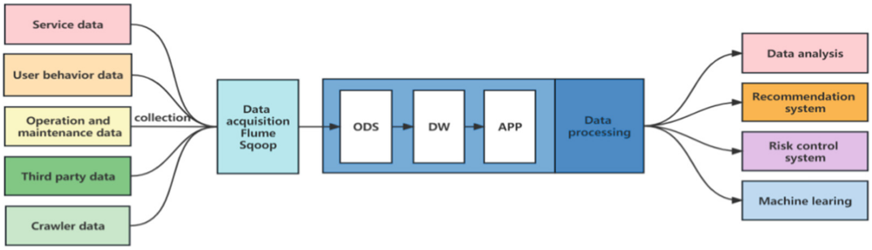


Fig. 2. Data platform layering [self-drawing]

Summary data layer, which uses "count + distinct" to summarize data and is implemented by "GROUP BY" statement; (5) APP/DA/ADS: Application layer data. It is final data which is beneficial to analysis panel formation [7, 8].

3 Panel Design

3.1 Hadoop and HDFS

Hadoop is a distributed system framework that allows users to use clusters for high-speed computing and storage without knowing underlying details. The core design of Hadoop framework is distributed file system (HDFS) and MapReduce. HDFS stores massive data, and MapReduce computes massive data.

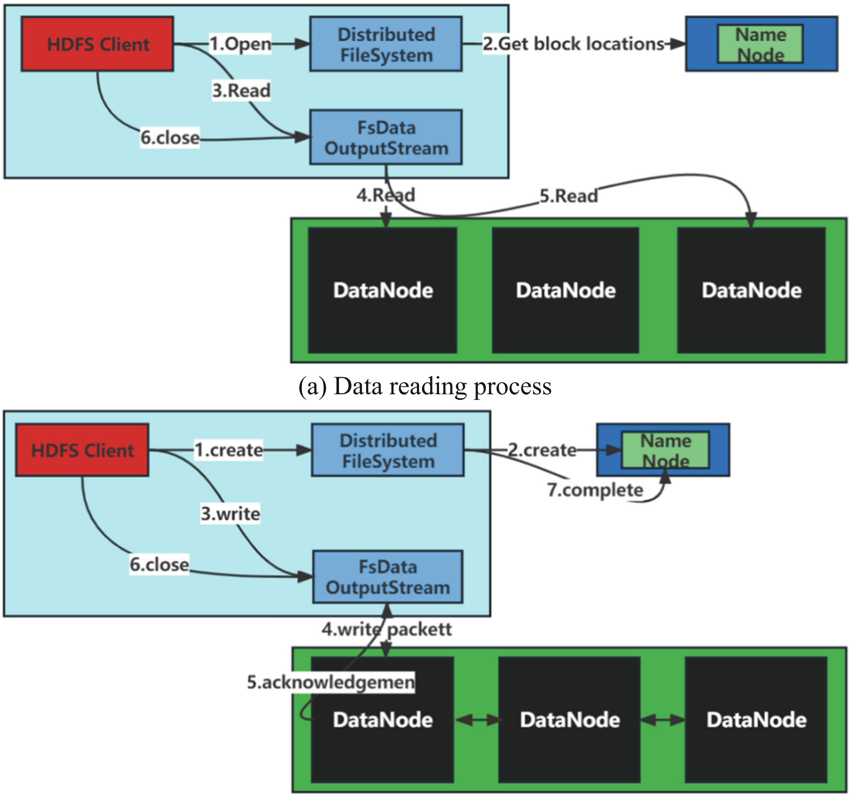
As shown in Fig. 3, HDFS has a primary node (Namenode) and multiple secondary nodes (Datanodes). The two nodes cooperate to read and write data in distributed file storage. HDFS mainly uses block storage. Data files are split into blocks and stored in different Datanodes [9, 10].

3.2 ODS Data and Data Conversion

ODS data sources mainly include customer service system, CRM database, student management system. For example, access consulting topics are mainly used to collect data on customer visits and consultations. The access data actually refers to number of customers visited. The data comes from customer service system database in MySQL, which contains two tables: "eb_chat_ems" and "web_chat_text_ems", which is shown in Fig. 4.

Registration and intention data come from CRM database, original data include two tables: "customer_relationship" and "customer_clue". There are six tables on dimension level, which are "customer", "employe", "scrm_department", "itcast_shcool", "itcast_subject", and "itcast_clazz". Attendance data comes from student management system, such as "tbh_student_signin_record", "tbh_class_time_table", "course_table_upload_detail", "class_studying_student_count", and "student_leave_apply".

In order to ensure data normalization from ODS layer to DWD layer, we need to convert the format and type of data. Specific conversion rules and examples are shown in Table 1.



(a) Data reading process

(b) Data writing process

Fig. 3. HDFS data reading/writing process [self-drawing]

Table 1. Data conversion example [self-drawing]

Original field	Conversion principle	Target field
msg_count(STRING)	From String to Int.	**msg_count(INT)**
create_time(STRING)	Cutting year	**yearinfo(TRING)**
	Cutting month	**monthinfo(STRING)**
	Cutting day	**dayinfo(STRING)**
	Cutting hour	**hourinfo(STRING)**

3.3 Data Export

Sqoop imports and exports data between relational database and HDFS. Sqoop program is converted to MapReduce program and submitted to YARN cluster for running [11, 12]. Data is exported to MySQL using an “sh” script, which is periodically triggered by “oozie” and executed after entire process is analyzed.

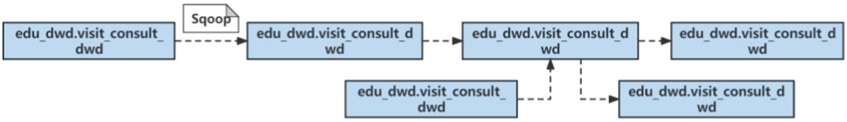

Sqoop
edu_dwd.visit_consult_dwd
edu_dwd.visit_consult_dwd
edu_dwd.visit_consult_dwd
edu_dwd.visit_consult_dwd
edu_dwd.visit_consult_dwd
edu_dwd.visit_consult_dwd

Fig. 4. Data stratification diagram [self-drawing]

Figure 5 shows that each panel requires multi-dimensional cleaning and transformation of original data table, such as user registration panel, user intention panel, student attendance panel.

As shown in Fig. 6, original data mainly comes from three business system databases. Service system contains data of users who visit and consult, RM system contains data of users, and attendance system contains data of student attendance.

The layer of registered users and intended users is mainly divided into five parts: ODS original data layer, DWD detailed data layer, DWM intermediate data layer, DWS summary data layer and APP application data layer. After data processing and conversion between each layer, data is finally transformed into visual data. In this paper, Sqoop is exported to local database MySQL, which provides effective support for FineBI visualization.

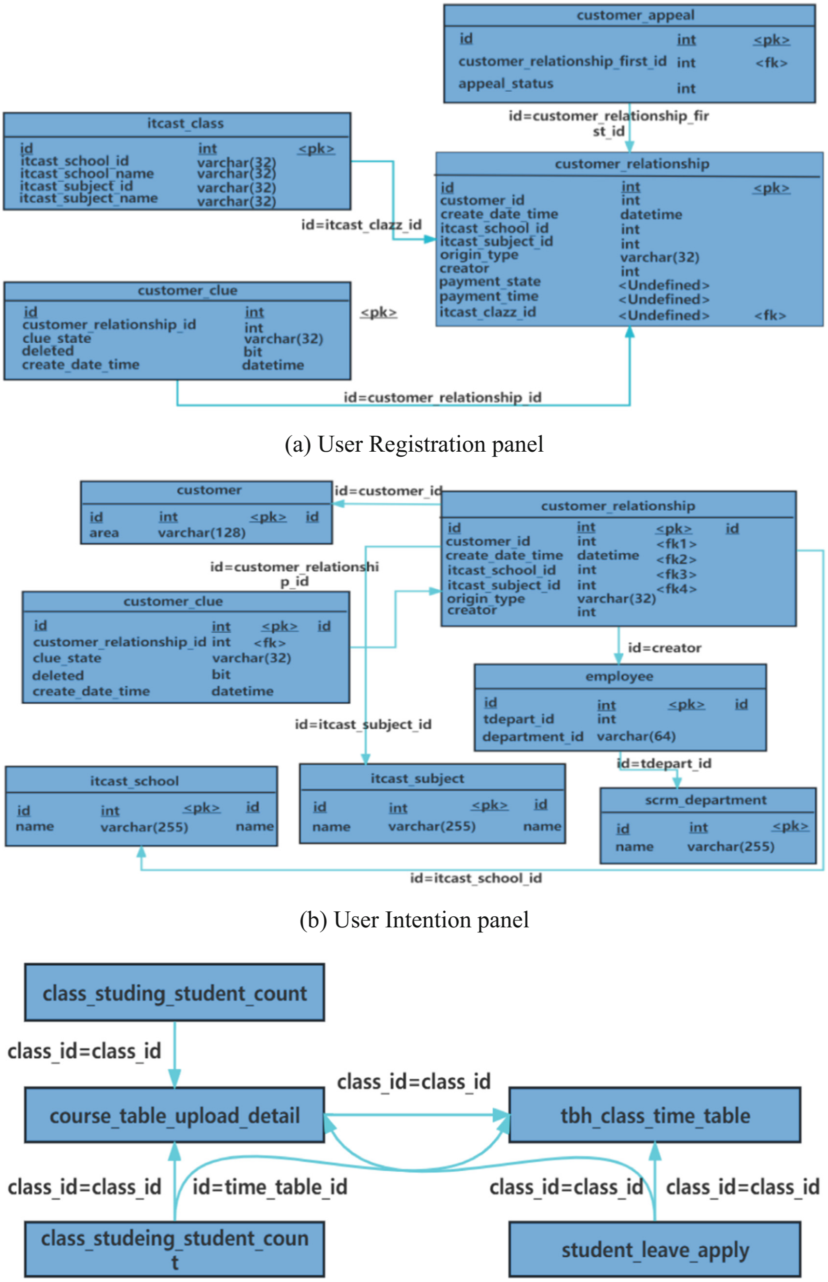
4 FineBI Visualization

4.1 Implementation of Access and Consultation Panel

Access and consultation panel mainly displays eight granularities, such as total number of customers visited daily, number of consulting users, daily independent visitors by region, trend of visiting customers, biaxial trend of customer inquiries and visitor inquiries per hour, proportion of daily source channel visits, and proportion of search source visits. The panel mainly displays the volume of visitors and inquiries through different interdimension, regional dimension and source channel dimension. As shown in Fig. 7, the changes of visitor volume and consultation volume data can be reflected through these comparisons, including how many visitor users can be retained to become consultation users, and which browsers are used to click and consult online education platform.

4.2 Implementation of User Registration Panel

User registration panel mainly displays seven granularities, which are total number of registrations, number of monthly registrations for each campus, number of monthly registrations for each subject, number of monthly online registrations, number of daily users, conversion rate of monthly online users, and Top number of subject registrations for each campus. As shown in Fig. 8, by displaying relevant information of registered users in seven granularities, we can see that there is a certain loss rate for prospective users change to registered users. Through this index, the loss of potential users can be effectively controlled and the number of registered users of platform can be increased.

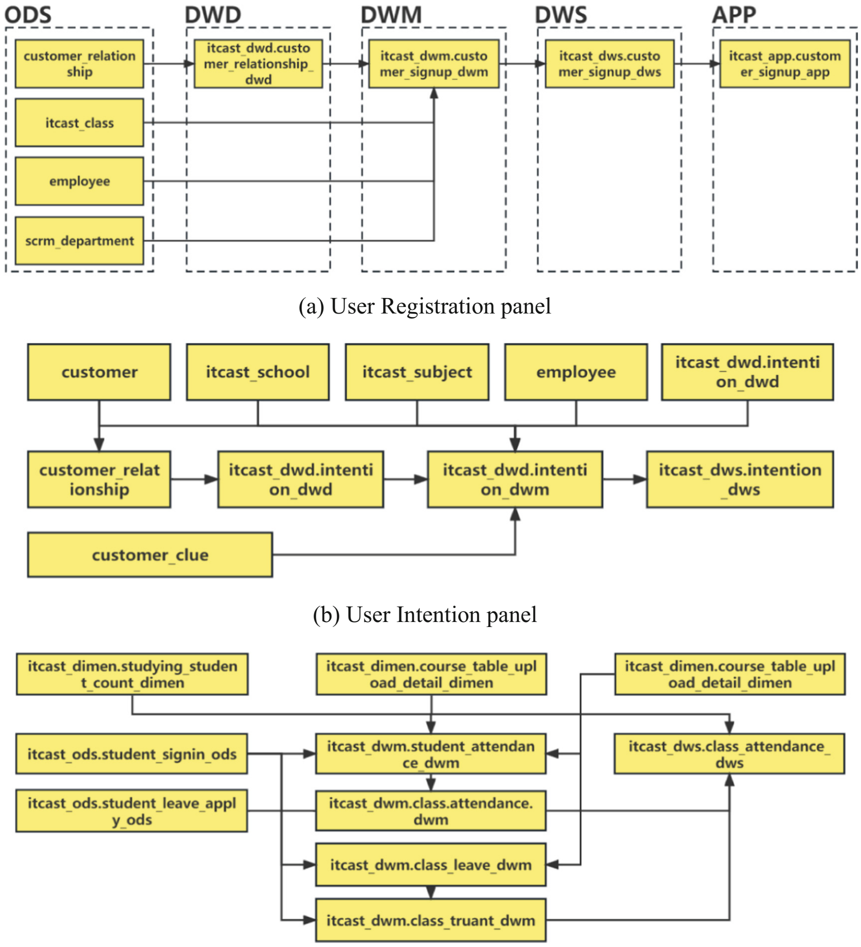


(a) User Registration panel

(b) User Intention panel

(c) Student attendance panel

Fig. 5. Source data structure of (a) User Registration panel; (b) User Intention panel; (c) Student attendance panel [self-drawing]


ODS
DWD
DWM
DWS
APP
customer_relationship
itcast_class
employee
scrm_department
itcast_dwd.customer_relationship_dwd
itcast_dwm.customer_signup_dwm
itcast_dws.customer_signup_dws
itcast_app.customer_signup_app
customer
itcast_school
itcast_subject
employee
itcast_dwd.intention_dwd
customer_relationship
itcast_dwd.intention_dwd
itcast_dwd.intention_dwm
itcast_dws.intention_dws
customer_clue
itcast_dimen.studying_student_count_dimen
itcast_dimen.course_table_upload_detail_dimen
itcast_dimen.course_table_upload_detail_dimen
itcast_ods.student_signin_ods
itcast_ods.student_leave_apply_ods
itcast_dwm.student_attendance_dwm
itcast_dwm.class.attendance.dwm
itcast_dwm.class_leave_dwm
itcast_dwm.class_truant_dwm
itcast_dws.class_attendance_dws

(a) User Registration panel

(b) User Intention panel

(c) Student attendance panel

Fig. 6. Hierarchical flow chart of (a) User Registration panel; (b) User Intention panel; (c) Student attendance panel [self-drawing]

4.3 Implementation of User Intention Panel

User intention panel mainly displays five granularities, such as total intention quantity, monthly intention quantity of each campus, monthly intention quantity of each discipline, online and offline intended users, and source channel statistics of intended users. Figure 9 shows that we can have a more systematic understanding of online education platform, as well as we can focus the improvement direction of sustainable development.

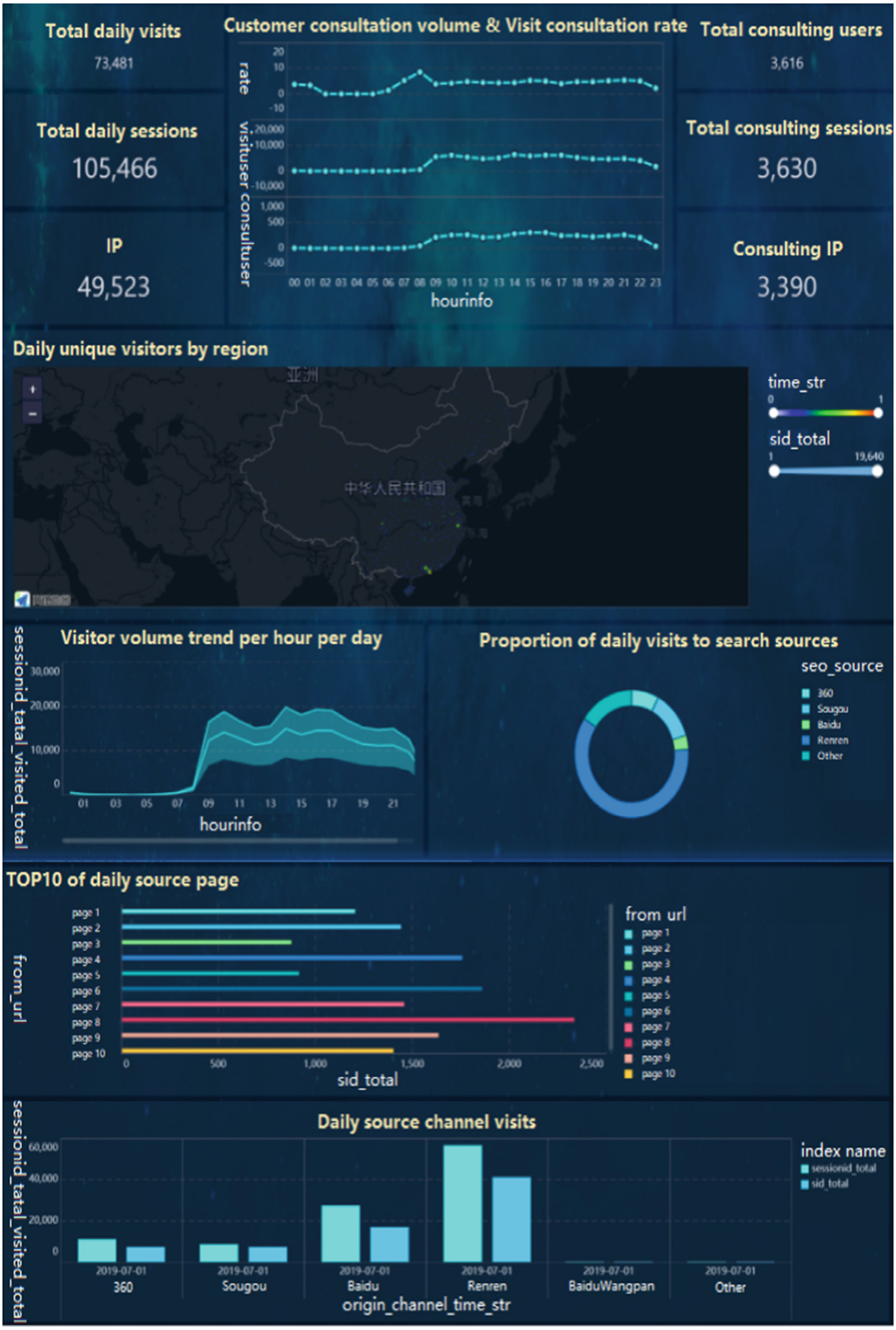
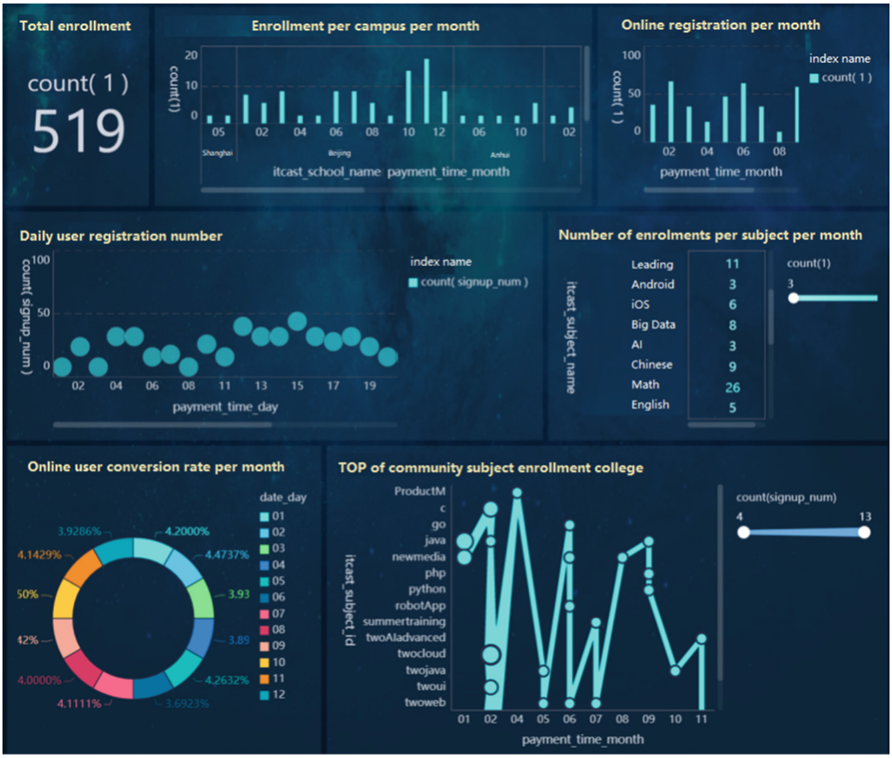


Fig. 7. Access and Consultation panel [self-drawing]

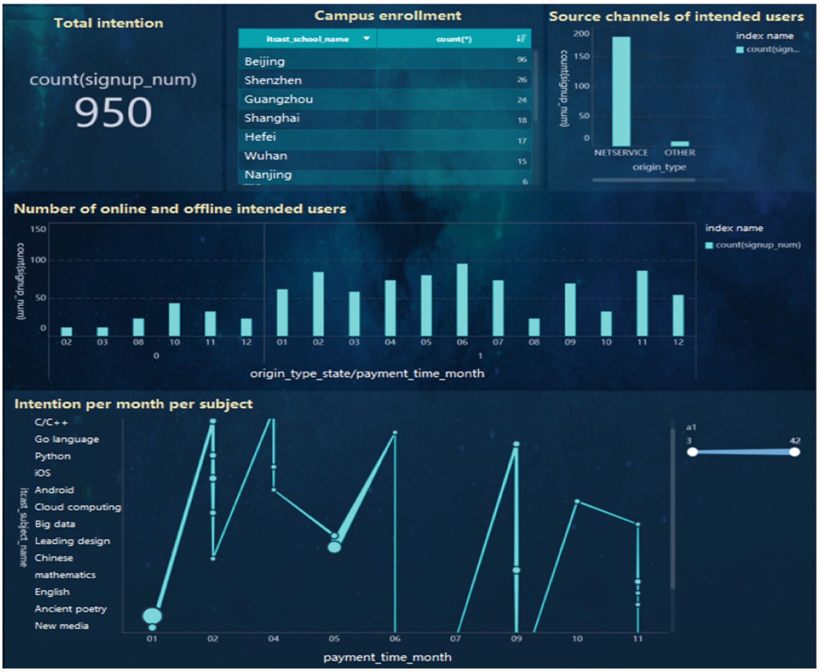


itcast_subject_name	
Leading	11
Android	3
iOS	6
Big Data	8
AI	3
Chinese	9
Math	26
English	5

Fig. 8. User registration panel [self-drawing]

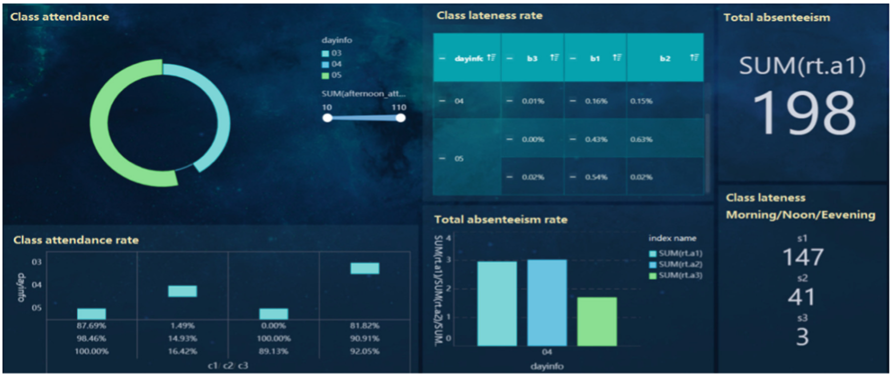
4.4 Implementation of Student Attendance Panel

Student attendance panel has six granularities, which are class attendance number, class attendance, total number of class absenteeism, total class absenteeism, class lateness number, and class lateness. Figure 10 shows that all these can well demonstrate learning status of each class, so as to ensure learning quality of each student. Only in this way can the platform develop efficiently.



itcast_school_name	count(*)
Beijing	96
Shenzhen	26
Guangzhou	24
Shanghai	18
Hefei	17
Wuhan	15
Nanjing	6

Fig. 9. User intention panel [self-drawing]



dayinfo	b3	b1	b2
04	0.01%	0.16%	0.15%
05	0.00%	0.43%	0.63%
	0.02%	0.54%	0.02%

Fig. 10. Student attendance panel [self-drawing]

5 Conclusion

This paper designs and analyzes online education platform based on Hadoop. It has configure HDFS, export data from MySQL through Sqoop, and build data platform based on YARN clusters. In this paper, ODS data is converted and written into DWD layer. Based on DWD layer, data is processed and written to DWM layer. We summarize DWM layer data and write results to DWS layer from time dimension, customer dimension and product attribute dimension. In addition, DWS layer data is analyzed and written to APP layer. The platform implements modules of access and consultation, intention and registration, and student attendance with APP layer data. Analysis panel results with FineBI visualization show that the number of intended customers converted to registered users and the loss rate. Data platform can help us effectively monitor the indicators of online education and ensure the quality of online education. It can be determined that it is essential for accurately increase user number of online education platform.

Acknowledgments. Special thanks to the following funds for their support: Key Research Project of Natural Science in Universities of Anhui Province (No. KJ2021A1160); Key Teaching Reform Research and Practice Project of Anhui Xinhua University (No. 2021jy006); Virtual Simulation Experiment Teaching Project of Anhui Provincial Quality Project (No. 2021xnfzxm046).

References

1. Ran Mu, Lin Chen, Xian Ye. Challenges and Countermeasures of Online Education Development under Situation of Internet+ [J]. Modern Business Industry, 2017, (35): 53–54.
2. Xing Ma, Nan Wang. Construction of University Teaching Quality Evaluation System based on Big Data [J]. Tsinghua University Education Research, 2018, 39(2): 38–43.
3. Jun Liu. Hadoop Big Data Processing [M]. People Publishing House, 2013, 44–75.
4. Yandong Li. Inspiration of Big Data to Solve Financing Difficulties of Small and Micro enterprises – A Case Study of ZestFinance Company [J]. Gansu Finance, 2016, (5): 28–30.
5. Yin Yang; Weiwei Gao. High Accuracy Recommendation Service of Hadoop Platform Applicable to Big Data of Online Education [J]. Machine Tool & Hydraulics, 2018, 46(24): 175–180.
6. Huihan Zheng. Research on Load Balancing Technology in Distributed Storage Based on HDFS [C]. Harbin Institute of Technology, 2016.
7. Yulong Zhang. Research on Innovation of Ideological and Political Education from Perspective of Big Data [D]. Harbin: Northeast Normal University, 2021.
8. Guofeng Ma, Jun Jiang. Research on Design of Project Objective Control Platform based on Big Data [J]. Science & Technology Management Research, 2018, 38(18): 216-221.
9. Mingyao Li. Personality Appeal and Resource Optimization of Online Education in era of Big Data [J]. Educational Theory and Practice, 2020, 40(04): 30–34.
10. Xiaoyuan Xu. Research on Interactive Teaching Path of Personalized Online Education based on Big Data Technology [J]. Computer and Information Technology, 2019, 27(04): 83–85+91.
11. Yanqiang Cui, Peipei Quan, Yelin Wu. Research on Realization Path of Teachers Precise Governance based on Big Data [J]. Journal of National Institute of Education Administration, 2018, (4): 11–17.
12. Jianjun Wang, Yingcheng Zhang. Research on Massive Structured Data Import in Universities based on Sqoop [J]. Wireless Internet Technology, 2018, 15(20): 52–53.