

Research on Centralized Data-Sharing Model Based on Master Data Management

Hong Dandan¹, Zhang Yajuan¹, Luo Junfeng^{1,*}, Liu Chen¹, Xu Mo¹, Suo Zhihai¹

¹Network and Information Center of Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China

Abstract. This research aims to analyse the data-sharing model widely used in information centres of Chinese universities. A university centralized data-sharing model (UCDSM) based on master data management system was proposed to provide unified data standards, exchange services, data-quality control and comprehensive data-sharing services for universities in big data environment. Then the core implementation methods of this data-sharing model were analysed and implemented. Finally, the application results show that the UCDSM can support large-scale data sharing for digital campus and all the data transmission processes can be monitored and managed under a unified management platform.

1 Introduction

With the rapid development of digital campus, many colleges and universities in China have built tremendous independent systems providing diverse services for different user groups in the campus [1][2].

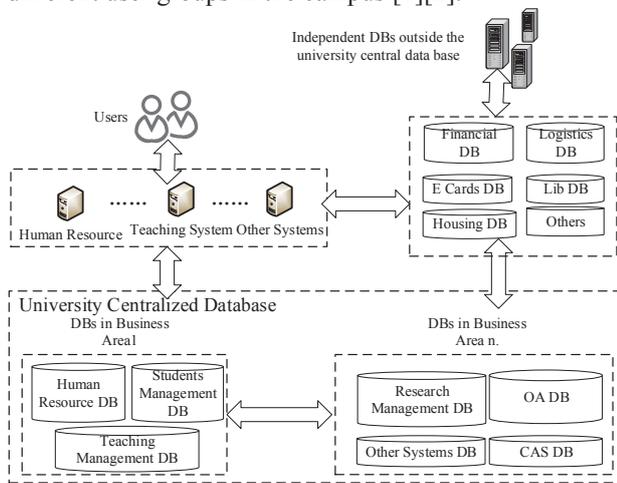


Figure 1. The traditional data sharing model in universities

As shown in figure 1, the campus database model commonly used in Chinese universities is consisted of two major parts. One is the university centralized database platform (UCDP) managed by the university network department, and the other is the independent databases managed by university business departments themselves [3].

On the university centralized database platform (UCDP), various independent databases belonging to different business systems are hosted by the network department centrally to reduce the pressure of database management in each business department. These model

brought benefits to the whole university information construction:

- The UCDP provides a unified hardware platform for business departments, so that they do not need to buy any database hardware servers or software.

- The UCDP provides central database management services to business systems including database operation monitoring, user accounts management, database security control, etc. So that the business departments can focus on the software development of their systems.

Due to these advantages of UCDP, the construction of digital campus has been highly improved in the last years [4]. But with the rapid developments of university information systems, the data and objects are stored in a large amount of repetitions in UCDP. This phenomenon brings storage pressure to UCDP, and affects the data sharing efficiency. Besides, the same type of data is usually stored in different encoding formats. If one business department wants to change some information of one object in its database or add new information to an object, the databases in other third-party systems associated with it cannot be updated synchronously. This imposes a major obstacle to data analysis in large data environments, because data analysts need to spend much time in cleaning and converting heterogeneous data in different formats [5].

To solve these problems, the work in [6] proposed to use MDM (Master Data Management) to resolve the issues of data duplication, inaccuracy and inconsistency. Baghi designed an architecture for master data application toward a decision model and it verified that the master data application architecture is better at the data quality aspect [7]. Gomed presented an approach to improving the data quality of a data warehouse based on MDM [8]. Jaksic analysed a common integration architecture using MDM [9]. These researches focused

* Corresponding author: luojf@xjtu.edu.cn

on building central master data management systems, which are crucial to improve the efficiency of data-sharing and data analysis.

Motivated by these related work, the university centralized data-sharing model (UCDSM) based on master data management system is proposed in this paper. The structure of our proposed UCDSM and the difference between UCDP and UCDSM will be discussed and analysed in section 2. Three major modules of our UCDSM including MDMS, master database and master data warehouse will also be explained in this section. Section 3 will describe the core steps of the data-sharing implementation in UCDSM. Section 4 will summarize this work and propose the future work.

2 Structure of UCDSM based on MDMS

The UCDSM proposed by this paper is constructed based on the UCDP described in section 1. To improve the efficiency of data exchange and data sharing, the UCDSM integrated MDMS, master database and data warehouse, therefore the UCDSM can integrate all the advantages of these three parts. The structure of the UCDSM is shown in figure 2.

As shown in figure 2, four main modules of UCDSM system includes master data management system (MDMS), master database, master data warehouse and the public database group. The four modules work together to implement the data sharing for services in university.

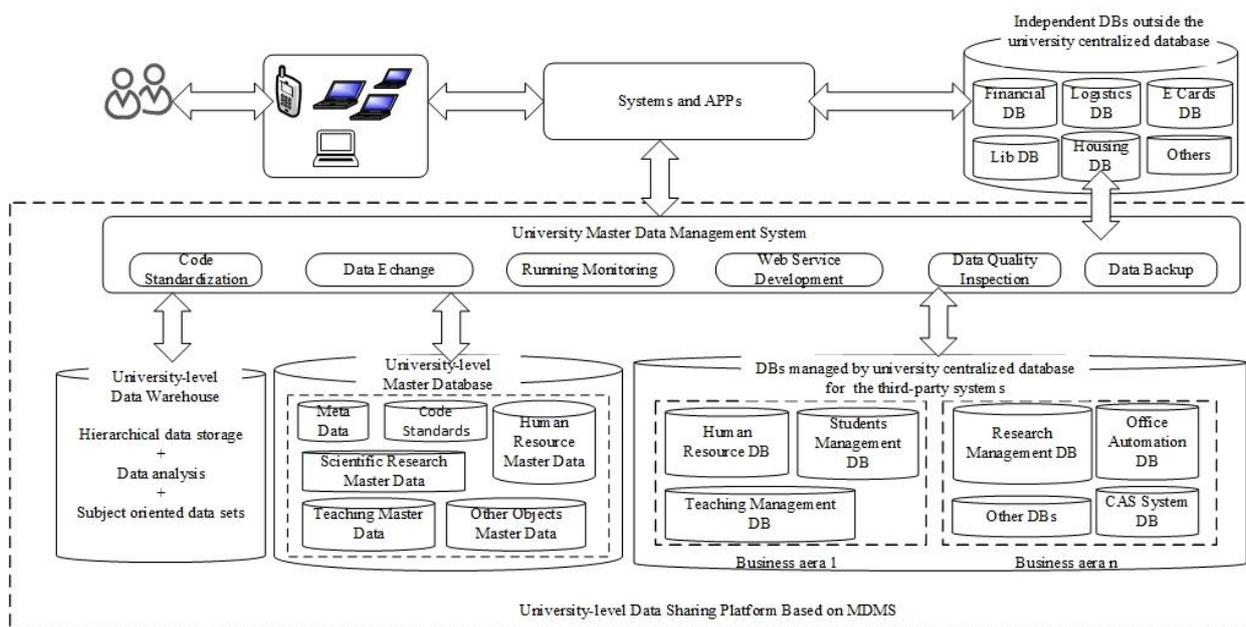


Figure 2. The university data sharing model based on MDMS

2.1 Master Data Management System

The Master data management system (MDMS) is the core part of the UCDSM. Its main functions are shown as follow.

Data standardization: it helps university establish, manage, publish and share its official data standards. The data collected from various data sources except master database must be formatted into one official standard of the university, therefore it can be shared to other systems requiring this data in digital campus. The detail process of data standardization is explained in section 3.2.

Data exchange: this module uses its ETL tools to implement the data exchange between the master database and the databases applied by business systems. The detail process is also described in section 3.2, along with data standardization.

Operation monitoring: this module includes system operation monitoring management, data integration monitoring, database operation monitoring and data flow

management. It helps the UCDSM managers to detect system exception and security threats in time.

Web service publishing: this module is responsible for interface development, management and authorization. According to the comprehensive data sharing scheme defined in section 3.2, it is necessary to develop web service interfaces for data sharing, for example the interfaces in JSON and XML formats. The systems that requiring data from UCDSM can get the shared data via the authorised interfaces.

Data quality inspection: this module is responsible for the data quality check, including metadata quality, master data quality. Whenever the abnormal data is found, such as inconsistent code, inconsistent data field in master database, then it gives alarms and rehabilitate suggestions to UCDSM manager in time.

Data backup: this module helps the master data warehouse to retain the historical data of university master data and reproduce the daily data, which provides an important method for data analysis in the time dimension.

2.2 Master Database

As shown in figure 2, the master database stores all kinds of master data managed by the MDMS, such as metadata, data encoding standards, the master data objects abstracted from each business database and the activity information of the objects. For example, master data of scientific research projects, human resource master data (HR master data), students master data, teaching master data and the other management data relating to them.

The shared data that required by related systems are provided in unified format by master database via the ETL tools of UCDSM.

2.3 Master Data Warehouse

The master data warehouse provides various data subsets for different analysis topics. For example, the research project dataset is used to analyse the scientific research achievements of the university. The teaching dataset is used to analyse the most popular lectures.

Meanwhile, under the help of OLAP (On-Line Analytical Processing), BI (Business Intelligence) and SDAS (Sale Decision Analysis System), the master data warehouse also supports hierarchical data storage and provides more valuable analysis results for university administrators.

3 Data Sharing Implementation based on MDMS

In this section, we define the university itself as the first level master object. The design pattern of its master data is discussed in section 3.1. The human resource master data is designed and discussed as its child object. Meanwhile, the core data-sharing implementation (Data standardization and data exchange) is also described in section 3.2.

3.1 Design Pattern of a University Master Data

Relying on the 2012 edition of the Chinese Ministry of education information standards and characteristics of universities, the university itself is defined as the first-level master object (Figure 3), all the other master data objects inherit it and extend its child objects and properties.

For example, the human resource (HR) master data, education resource master data and the other business master data as shown in figure 3 are defined as the second-level university master data.

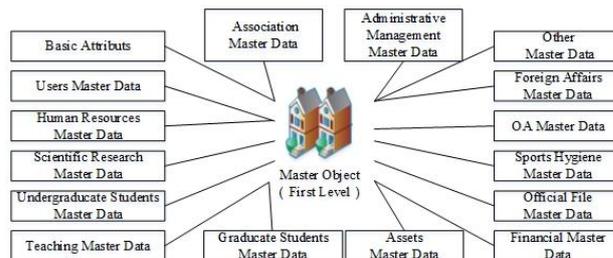


Figure 3. Design Pattern of the First Level Master Data (Only the first and second levels are displayed)

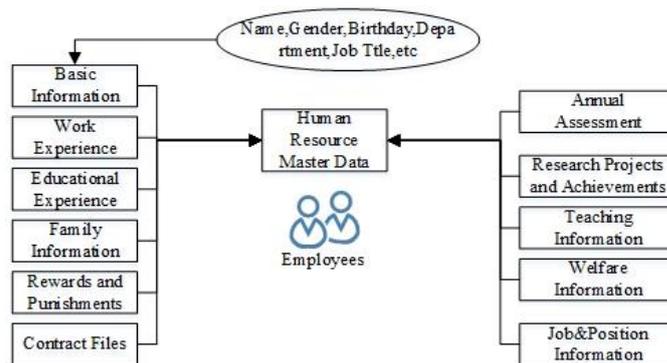


Figure 4. Design Pattern of the HR Master Data

As shown in figure 4, the HR master data model of the university is consisted of 11 child objects currently, namely basic information, work experience, education experience, family information, rewards and punishments, contract files, annual assessment, research projects and achievements, teaching information, welfare information, job&position information. Besides them, the model can be enriched synchronously with the development of the master data construction.

3.2 Implementation of Data Standardization and Exchange

As mentioned in section 2, data standardization and data exchange are the core steps of the whole data sharing process, which consists of three steps: data integration, data cleaning and data sharing (figure 5). These steps work together to implement the data sharing of the UCDSM in high efficiency and help universities conduct more detailed and standardized data analysis in big data environment.

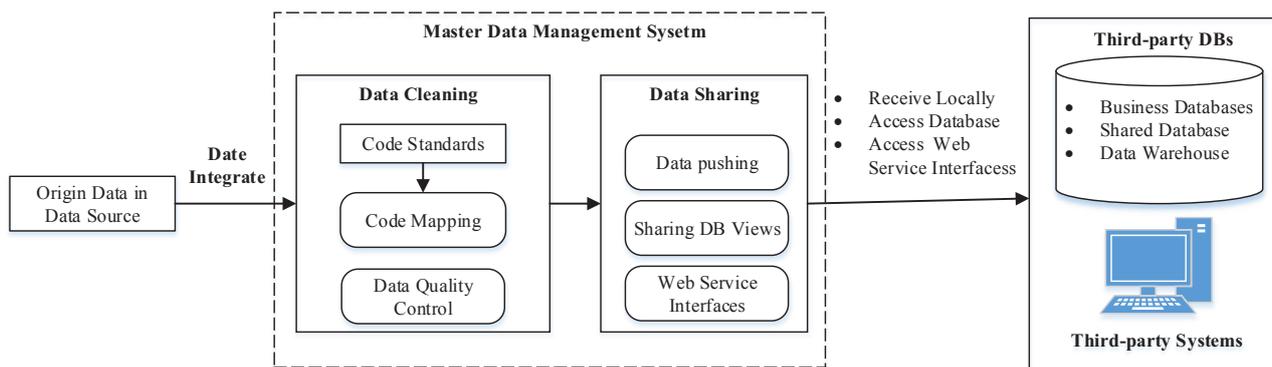


Figure 5. Data Standardization and Exchange Process

3.2.1 Data Integration

MDMS firstly finds information about the data source from the meta database stored in the master database. Then MDMS sends an access request to the data source and extract the data it needs.

To ensure data extraction efficiency, there are two approaches to extract data from databases as follows.

- The whole data extraction

MDMS extracts all the data it needs from the data sources into the temporary area of the master database. The advantage of this approach is that the retrieved data is a whole copy of the data stored in its original database, so it can keep in line with it in three aspects, namely integrity, quantity and timeliness.

The disadvantage is that if the amount of data is large (for example, all the employees' photos in Oracle blob format), the data integration process will spend much time for data extraction. Besides, if the data transfer process is accidentally terminated for some reason, the MDMS must restart this data extraction procedure no matter if the same data has been already extracted.

- The incremental data extraction

This approach only extracts the data that has been changed in its source system after the last extraction. The advantage of this approach is that it can greatly reduce the amount of data to extract and improve the efficiency of data extraction. Meanwhile, the data timeliness and data integrity can also be guaranteed.

Based on the above analysis, the second extraction (incremental data extraction) approach is adopted in UCDSM proposed in this paper.

3.2.2 Data Cleaning

Data cleaning consists of two parts: data quality control and code mapping.

- Data quality control

Data quality control is responsible for two tasks. One is the data format conversion. The other is data quality monitoring. Data format conversion aims to unify different formats of the same data into standard data formats. For example, convert different types of time formats (e.g., 2017-9-17 12:30, 2017/9/17 12:30:00, etc.) into the unified time format (e.g., 2017-09-17 12:30:45).

Data quality monitoring aims to check the contents, standards and code formats of the retrieved data. If these indicators do not meet the requirements of the MDMS, alarms will be shown in the MDMS.

Database1		Code Mapping Table in MDMS		
Gender	Code	Gender	Code	Database name
Male	M	M	01	Database1
Female	F	F	02	Database1
Unknown	U	U	03	Database1
Database2		1	01	Database2
Gender	Code	2	02	Database2
Male	1	3	03	Database2
Female	2	a	01	Database3
Unknown	3	b	02	Database3
Database3		c	03	Database3
Gender	Code			
Male	a			
Female	b			
Unknown	c			

Figure 6. Code Mapping Table in MDMS

- Data code mapping

The data code mapping creates a code mapping table in the master database and establishes the mapping relationships between the code stored in the original database and the master code defined in the master database.

As shown in figure 6, the gender code of an employee, in database1 (DB1), "M, F, U" are encoded for male, female, and uncertainty respectively. In database2 (DB2), "1" is for "male", "2" is for "female", and "3" is for uncertainty. Under this scenario, the MDMS needs to unify these data encoding rules according to the same standard. For example, all the codes extracted from different systems for "male", "female" and "uncertainty" are re-encoded as "01", "02", and "03" respectively.

3.2.3 Data Sharing

Data Sharing model provides three approaches to help the third-party systems get the standard data from university master database.

- Data pushing

The third-party system provides the JDBC remote connection authority to the MDMS. Then the MDMS uses a data-exchange tool (e.g., Oracle ODI tool) to push

* Corresponding author: luojif@xjtu.edu.cn

the standard data directly into the database defined in the JDBC URL.

- Sharing DB views/tables

In this approach, the MDMS creates a data-sharing view (or table) in the master database, then the shared view can be authorized to multiple systems. And the third-party systems decide if the shared data should be copied and stored in its own database.

- Web service interfaces

In this approach, MDMS creates web service interfaces (JSON, XML) for data sharing. If the third-party systems need to get these data, MDMS authorizes the third-party systems to access the interfaces according to the roles of the third-party systems.

All the advantages and disadvantages of these three approaches are compared and summarized in the following table 1.

Table 1. Advantages and disadvantages of the three data-sharing methods in UCDSM.

Method	Advantages	Disadvantages
Data pushing	<ul style="list-style-type: none"> • The shared data are stored in the third-party system, and the third-party system does not need to connect to the master database frequently, so it can reduce the access pressure of the master database. • MDMS is responsible for data synchronization at regular time. The third-party systems need not to maintain the shared data, so the working pressure is reduced. 	<ul style="list-style-type: none"> • The same type of shared data are stored repeatedly in many third-party systems within the campus. • MDMS must create several data-pushing procedures (for example, oracle ODI procedures) to update the shared data stored in various third-party databases.
Sharing DB views(tables)	<ul style="list-style-type: none"> • MDMS does not need to create data-exchange interfaces, and all the granted third-party systems can access to this shared view directly to get their requiring data. • The same shared data are only stored in one copy in the university master database. The third-party systems decide whether it should be copied locally. 	<ul style="list-style-type: none"> • The number of the third-party systems requiring to access the master database directly will continue increasing. This brings pressure to the master database. • There are too many users accessing the same shared view, which brings risk to system security and data security of the master database.
Web service interfaces	<ul style="list-style-type: none"> • All the shared data are encapsulated from the web service interfaces. The third-party system can not access to the master database directly. • The data in the interface are strictly consistent with the master database 	<ul style="list-style-type: none"> • There are too many users accessing the same shared web service interfaces, which brings system access pressure to master database.

Through these comparisons, a comprehensive scheme is proposed as follows.

- If the thirty-party systems have highly frequent access which will brings access pressure to the master database, the first way is highly recommended.
- If the shared data are changed frequently, and there are less third-party systems requiring these data, the third way is highly recommended.
- The second way is not recommended. Because the DBA of the master database in UCDSM must manage a large amount of DB accounts, the relationships between the accounts and the DB views. This will bring pressures to DBA and database management. Meanwhile, the data security is also threatened if several accounts access to the same shared DB view or DB table.

4 Application and Conclusion

The university data sharing model based on MDMS designed in this paper has already been applied in Xi'an Jiao Tong University (XJTU).

As shown in figure 7(a), 2332981 records have already been stored in the university master database. And 468991 records are stored as payment master data whose parent is the finance master data. 267735 records are stored as curriculum schedule master data whose parent is the teaching master data.

Figure 7(b) shows that there are more than 14 types of code standards managed by MDMS to support the code mapping and data exchanging between the data sources and the third-party systems. The public code

standards, teaching affairs code standards and scientific research code standards are the top 3 standards in the MDMS (statistics based on code quantity).

Top10 Ranking List of Master Data Quantity (Total 2332981)

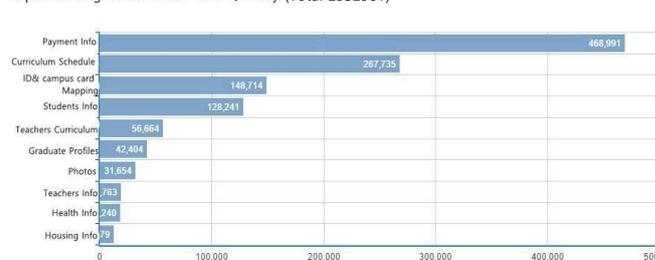


Figure 7 (a)

Statistical List of Code Standards Quantity

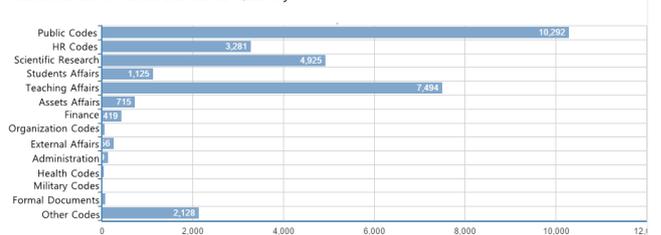


Figure 7 (b)

Figure 7(c) shows that there are more than 25 data sharing interfaces integrated into the MDMS, which are distributed in various systems, such as teaching management system, organization management system, human resource management system and so on.

Figure 7(d) shows that there are more than 373 thousands data records shared among 5 systems. For

example, the all-in-one card system contributes the most shared master data in the university campus.

Top10 Ranking List of Integrated Interface Quantity

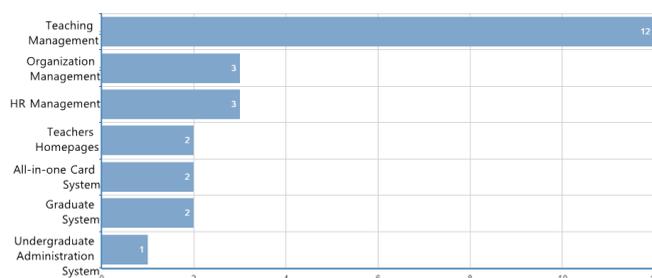


Figure 7 (c)

Data Activity Statistics of Business Systems

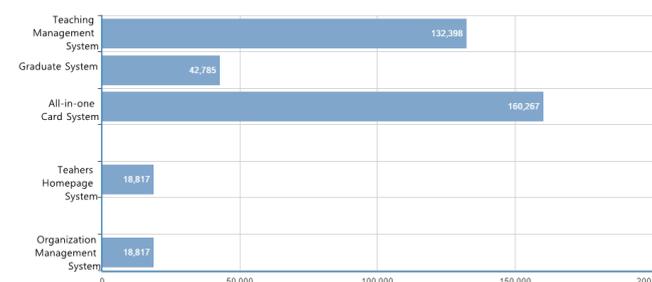


Figure 7 (d)

Figure 7. The application results of UCDSM in XJTU

In summary, the application result shows that this model can satisfy the requirements of large-scale data sharing management in digital campus. The joint analysis of different data in XJTU has been greatly improved with the help of unified data standard, unified data quality control in the big data environment.

In the future, more effort will be devoted to research on data mining of multiple data in university campus, e.g., the analysis of education data [10], research on providing accurate service based on users' behaviour [11] [12]. Meanwhile, the university master data warehouse will also be concerned to support more official and authoritative decision analysis [13] [14].

References

1. Li M J. The Digital Campus Information System Integration Research Based on SOA[C]// International Conference on Intelligent Computation Technology and Automation. IEEE, 2015:781-784.
2. Zhu R, Chen Y, Zhuang C. Experimental Center Construction of Teaching Management Platform under the Background of Digital Campus[C]// Eighth International Conference on Measuring Technology and Mechatronics Automation. IEEE, 2016:286-289.
3. Yang M, Zhang X. The research of resource sharing platform for universities using virtual database[C]// Advanced Research and Technology in Industry Applications. IEEE, 2014:690-693.
4. Zhu R, Chen Y, Zhuang C. Experimental Center Construction of Teaching Management Platform under the Background of Digital Campus[C]//

5. Xu Q, Dazhi. The application of heterogeneous database general query in digital campus[C]// Cross Strait Quad-Regional Radio Science and Wireless Technology Conference. IEEE, 2011:1193-1195.
6. Haneem F, Ali R, Kama N, et al. Resolving data duplication, inaccuracy and inconsistency issues using Master Data Management[C]// International Conference on Research and Innovation in Information Systems. 2017:1-6.
7. Baghi E, Schlosser S, Ebner V, et al. Toward a Decision Model for Master Data Application Architecture[J]. 2014:3827-3836.
8. Gomed E, Barros R M. Master Data Management and Data Warehouse: An architectural approach for improved decision-making[C]// Information Systems and Technologies. IEEE, 2013:1-4.
9. Jaksic D, Jovanovic V, Posic P. Integrating evolving MDM and EDW systems by data vault based system catalog[J]. Proceedings of Jubilee International Convention on Information & Communication Technology Electronics & Microelectronics, 2017:1401-1406.
10. Shu J, Wang X, Wang L, et al. Exploration on college education big data open service platform[C]// IEEE, International Conference on Cloud Computing and Big Data Analysis. IEEE, 2017:161-165.
11. Guivarch V, Paz J F D, Villarrubia G, et al. Hybrid system to analyze user's behaviour[C]// Computational Intelligence. IEEE, 2017:1-5.
12. Yu C. Sale Decision Analysis System Design Based on Data Warehouse[C]// International Conference on Intelligent Systems Design & Engineering Applications. IEEE Computer Society, 2015:712-718.
13. Khanna L, Singh S N, Alam M. Educational data mining and its role in determining factors affecting students academic performance: A systematic review[C]// India International Conference on Information Processing. 2016:1-7.
14. Singh R P, Singh K. Design and Research of Data Analysis System for Student Education Improvement (Case Study: Student Progression System in University)[C]// International Conference on Micro-Electronics and Telecommunication Engineering. IEEE, 2017.