Design and Implementation of Basic Data Platform for University Library

Zehua Fan1, Jiao Song1*

1Library, Huazhong University of Science and Technology, Wuhan, Hubei, 430074, China
*Corresponding author’s e-mail: songjiao@hust.edu.cn

Abstract. Considering the problems existing in the data management and use of university library, this paper puts forward the concept of constructing "basic data platform", which can provide the guarantee for the production and business data of library and university. The system includes data acquisition system, library data exchange system, data access API interface, cold data transfer system, data visualization and management platform, data analysis and mining modules, forming a complete system of data storage, management and use.

1. Introduction
The "Public Library Law of the People's Republic of China"[1] came into effect on January 1, 2018. The eighth item of the Public Library Law is "The state encourages and supports the use of science and technology in the construction, management and service of public libraries, and promotes the application of modern information technology and communication technology to improve the service efficiency of public libraries." To promote the development of library from digitization to intelligence, National Library puts forward the construction of "National Wisdom Library System"[2]. As an important part of the library system, university library has been promoting the information construction in recent years. The lack of unified planning and design of data in library leads to the independence of systems and the formation of information island. Taking the library of Huazhong University of Science and Technology as an example, this paper constructs a secure and extensible basic data platform for data storage, management and transfer, which supports the acquisition, cleaning, storage and use of data, and provides a guarantee for the production and business data of libraries and university.

2. The necessity of the construction of basic data platform in University Library
Library has always been an active advocate of new technology, but also the beneficiary of technological progress. In the process of promoting information construction, all kinds of information systems are gradually increasing, their construction methods and needs are different, so there is no channel for interconnection between those systems, forming an information island.

2.1. Need for data statistics
The data of each business system is increasing day by day, involving all aspects of library business, which is an important guarantee for library operation. For data statistics and analysis, some systems can operate the database directly, most closed systems can only use simulated page access, which requires the librarian have rich information literacy and professional knowledge background. It's even more difficult to joint multiple system queries.
2.2. Need for data exchange with the outside world
In order to promote the construction of university informatization, the university has established a public data platform and explored the "One Table" project [3]. It needs to be discussed on how to synchronize the reader data from the university to all the business systems of the library, and how to push the authoritative data of the library to the public data platform of the university.

2.3. Need for personalized data analysis for readers
The essence of library is the intermediary of knowledge and information service. Under the condition of the network environment and user demand, the library needs to speed up the transformation from resource ability to new service ability [4]. It is necessary to effectively integrate the library data to better convey the resource utilization service to the readers.

Owing to the above reasons, the library needs to establish a secure and extensible basic data platform for storing, managing and invoking data, supporting the acquisition, cleaning, storage and use of data, and providing security for the production and business data of libraries and university. This paper introduces a practical experience in the construction of basic data platform in library of Huazhong University of Science and Technology.

3. Design of basic data platform system for University Library
Taking the practical needs of the library of Huazhong University of Science and Technology as an example, this paper designs a set of basic data platform of the library in view of the needs of data access, synchronization and statistical analysis encountered by the library in its practical work.

The overall framework of the system platform is shown in figure 1. The whole platform takes the library data warehouse as the core, and consists of the following seven modules:

- Library data warehouse, which is used to store all kinds of business data, resource metadata, reader data and so on.
- Data acquisition system, which collects, converts and processes data in our library and stores them in the data warehouse.
Data exchange system, through which the data exchange between the library and the university can be realized.

Data access API interface, which provides the API interface of data access to the third-party system, and realizes the sharing of data among multiple systems.

Cold data transfer system, which transfers the cold data to the existing big data system.

Data visualization and management platform, which is used for statistical analysis of data and the statistical results are displayed on multiple terminals through visual diagrams.

Data analysis and mining. Through deep learning and mining of data in the platform, various useful data models are extracted to realize the portrait based on readers or resources.

4. Realization of basic data platform system in University Library

The system includes several subsystems, which perform their own functions and cooperate with each other to realize the functions of the basic platform.

4.1. Data acquisition system

The data acquisition system is mainly responsible for collecting the original data from various existing systems in the library into the basic data platform. The data can be divided into three categories: service data, resource metadata and management data. This system needs to deal with the problems such as different ways of data storage, different updating frequency and different fields in different business systems.

Business systems’ data is saved in different ways. Most systems store data in the form of databases, including the common traditional relational databases, such as Oracle, MySQL, SQLServer and so on. Some systems are non-relational databases, such as MongoDB. A few closed systems do not provide access to the database, the collection of these data needs to be carried out by simulating web page access. In addition, there are some manually recorded data in the form of EXCEL in library. Thus, the acquisition system is required to be compatible and suitable.

The update frequency is different, so the acquisition system is required to set the frequency separately according to the characteristics of each system. For some data with high frequency of updates, synchronization time should be set to be as short as possible to improve the real-time performance, such as seat reservation data. For some infrequent data, a little longer synchronization time can be set to reduce unnecessary system consumption, such as degree thesis metadata.

The problem of field difference is the biggest problem in the process of data acquisition. The data of each system is irregular, repetitive or missing, which needs to be filtered and cleaned before to be imported into the system.

4.2. Data exchange system

The data exchange system is mainly responsible for the data exchange between the library, the university and the third-party system. At present, there are many kinds of data to be exchanged. But the field definition of the same property value is different in each system, the data need to make a intermediate conversion before synchronizing.

One of the major tasks of this system is the maintenance of reader data. The architecture of the reader data synchronization system is as follows:
At the bottom of the system, the one card system and the university basic database system are bases of the architecture diagram. The source data of the two systems are synchronized in the form of "one card third party application API" interface and intermediate table respectively. Among them, the one-card system has the information about card number, card status, card validity and other fields; the basic database of the university, which integrates the undergraduate data of the academic department, the graduate data of the research department, the teacher data of the personnel department, includes the name, gender, department, major, status and so on. When readers enter the library and use card to borrow or return books, card information and university status information are needed. The library effectively integrates the fields of the two data sources and constructs a complete reader identity information, which is stored in the basic data platform of the library.

Considering the loss of card, the graduation of students, the departure of teachers and students, the information of teachers and students is changing anytime and anywhere. If the full amount of data is obtained regularly, the system will be inefficient, burdensome and delayed for a long time. Therefore, a data update method is provided: the API interface, which allows the two data sources to actively call API to update sporadic data. After a third-party system calls the API interface, the system sends the data update requirements to the background through the message queuing system, effectively decoupling the system load.

Reader data processing system [5] is the core part of the whole system, through the interface or intermediate table to maintain the connection with each data source, at the same time will respond to the message queue sent to update the data request, timely to obtain the latest data. All the data obtained by the system are updated to the reader table in the basic database of the library in real time, and updated to the relevant business subsystems of the library at the same time.

Reader data is the basic data of the library, a large number of business systems need reader information to ensure the normal development of the business, such as library integrated management system, access control system, institutional knowledge base system and seat reservation system and so on. However, the field definitions of reader data in these systems are different. When synchronizing the reader tables in the basic database to different business systems, it is necessary to carry out some intermediate conversion according to local conditions.

4.3. General form filling system:
There are a lot of business forms involved in the daily business of the library. The office needs to operate at three levels and five steps before getting a corresponding report, which is time-consuming, laborious
and prone to errors and omissions. The contents of different business forms are repeated and filled in many times, which is not convenient for the unified preservation and management of data.

In order to solve the problems, a general form filling system is provided in the basic data platform. The staff member can create some new templates with the content that needs to be filled, and then inform the person that is concerned to enter the system and fill it out.

The general form filling system has the following advantages:
- supporting multiple people to fill in online at the same time;
- supporting the historical data selection function, solving the problem of repeated filling of data;
- supporting data sharing between different tables;
- supporting to keep all historical data.

4.4. Data access API interface

Many business systems need to access some public basic data, including the various business systems built by the library, the systems of cooperative manufacturers and the systems of the relevant units of the university. In order to facilitate the access of all kinds of systems to the library in a timely and convenient manner, a set of data access API interface is designed:
- The reader information query interface, can query the reader's identity data according to the reader's university number, name or physical card number.
- The bibliographic query interface, can query book detailed metadata information according to book B record or I record.
- The reader loan record query interface, can query the reader's book borrowing record according to the reader's academic number or name.
- The reader access control data query interface, can query the reader's entry records according to the reader's university number or name.
- The reader operation record query interface, can query the reader's comprehensive information in the library according to the reader's university number or name.

The system provides the authority management function of the account, only after the administrator has approved, the third-party system can call the data interface. In addition, the system also integrates the API interface call number statistics, access time limit and other interface control functions.

4.5. Cold data transfer subsystem

The basic data platform stores the real-time data generated during the operation of each business system. After a period of operation, the access frequency of the "real-time" data will be met gradually lower, converted into cold data. If we ignore the cold data, that will be very large over the years and impact the whole platform system’s performance. But if we clean them up directly, then the original intention of the construction of the basic data platform will be lost. Thus, it is a good idea to transfer cold data to another platform.

Library big data storage system, completed in 2016, is responsible for unified storage, unified organization and unified management of structured and unstructured data in the library, which has become the best choice for receiving cold data in the basic data platform.

Therefore, the design of cold data transfer subsystem is responsible for docking the basic data platform with big data storage system. Big data storage system is based on Hadoop, the system uses impala docking to make sure cold data can be extracted and transferred to big data system. The system supports the personalized configuration of transfer cycle and transfer content according to the characteristics of each business data.

4.6. Data analysis and mining

The construction of the basic data platform completes the data collection and integration, on the basis of which the data analysis and mining system is established, the association and fusion of all kinds of business data are realized, and the fine-granularity portraits based on readers and resources are constructed, so as to effectively improve the individualized service level of the library.
The personalized service ability of readers in library is generally weak, which is due to the independence of each business system and its data, and the lack of a platform for linkage and overall planning. When making the annual report or reader recommendation, the staff can only extract a certain aspect information in each business system, then summarize the statistics. This information has no internal relationship and the analysis results have a certain degree of one-sidedness and lag.

On the basis of the solid data, the data analysis and mining system adopts the way of flow processing to process and analyse the data in real time. For example, when a reader searches for books on a library website, the system will calculate the books he may be interested in according to the reader's historical borrowing records, recent search records, and personal information such as the reader's department and major, and then dynamically display them on the search interface to improve the reader's experience. Further, the correlation analysis between access control data and readers' borrowing records is carried out. At the moment when readers swiping their cards into the library, book information is actively pushed through WeChat or SMS.

4.7. Data visualization and management platform

The data visualization and management platform is responsible for the system configuration and monitoring of the whole platform. It can uniformly view the running status of each subsystem and automatically alarm when abnormal occurs. The administrator can quickly locate and solve the problems through the system log files.

The system provides the data statistics function, which can provide the related statistical report and query function for the data of each business system. The daily statistical needs of the library are very complex, so the system provides the statistical configuration function. The administrator can configure according to the concerns and needs of each department, or even input a SQL statement directly to complete the particularly complex data statistics.

After the statistical analysis, the results can be displayed publicly, so the data visualization platform is also responsible for the docking with the library big-screen management system to display the statistical data in a visual way. In addition to the big-screen display, the data visualization platform also supports other terminal platforms in the library, including APP, WeChat, PC and so on. For different display terminals, the system supports customizing the display content of each terminal and supporting what you see is what you get to adjust the layout of each data item.

5. Conclusion

Through the construction of the basic data platform, the problem of "information isolated island" is solved, the interworking and interconnection of heterogeneous systems is convenient. The effective synchronization of data among multiple systems is realized, which is consistent with the authoritative data source. In the future, the new business system should be accessed according to the standard system, thus the management can be expanded.

References

[1] Government. (2017) Law of the people's Republic of China on Public Library [EB/OL]. Http://www.gov.cn/xinwen/2017-11/05/content_5237326.htm.
[2] Rao, Q. (2021) National Smart Library System: Opening a New Chapter for the Transformation toward Smart Libraries. J. Journal of the Chinese Library., 47 (01): 4-14.
[3] Liu, Y.Q., Mao, W.H., Xiong, Y., Li, K., Liu, X.L. (2021) Practice of Data Management of College Teachers based on "One Table" Project. J. Educational Informatization in China.,2021(03): 48-51.
[4] Chu, J.L., Zhao, Y. (2019) Library Capacity Transformation from Resource-based to Service-based Model. J. Library and Information work., 63 (01): 11-17.
[5] Song, J., Fan, Z.H. (2021) Design and Implementation of Reader Information Synchronization System in University Library: Case study of Huazhong University of Science and Technology. J. Journal of Sichuan Library., 2021 (01): 36-40.