Data Quality Analysis and Integration in University Information Systems

Yong-jun LI¹ and Liang ZHOU²,a*

¹Personnel Department of Shanghai Ocean University, Shanghai, 201306, China;
²Network and Information Center of Shanghai Municipal Educational Exams Authority, Shanghai, 200433, China

*izhou@shmeea.edu.cn
*a*Corresponding author

Keywords: Data quality, Data integration, Information systems, Data warehouse

Abstract. University information systems (UIS) are an important part of digital campus, which makes the management and decision-making more convenience. With mass of data emergence, people have to think of data quality and data integration due to the issues created by the gap between the legacy system and the current systems. The effect of UIS application is directly related to data quality. Furthermore, there exists data isolation between different systems, which leads to data inconsistency and difficulties in sharing information. Consequently, we need to analyze the existing data quality and propose relevant basic concepts. This paper takes Shanghai Ocean University (SOU) digital campus construction as an example to show the actual application of UIS data integration.

Introduction

With the development of digital campus, there are different technology, system framework and information standard due to historical reasons. Therefore, it is important to establish a comprehensive service platform of university, such as "service platform"[1]. It can make data integration of the existing information system, fully implement of digital campus construction, and guarantee the efficient and orderly data, service, authentication and other aspects of the management. However, it is needed for data quality analysis, analysis and data quality of data integration organically before the construction of the service platform.

Data Quality Analysis

Data quality analysis includes (1) the definition of data quality, (2) domain analysis and filter analysis, and (3) analysis of ETL based on data warehouse [2].

Definition of Data Quality

The quality of data has different definitions in different industries and organizations, but generally it can be used to characterize with the following 6 properties, which is illustrated with relevant examples.

Correctness: response data shows the objective world, and values should be in the corresponding data domain. For example, the doctoral or master's degree professional code must conform to the subset of the standard code for the Ministry of Education.

Accuracy: on the basis of correctness, accuracy reflects how data matches with the objective world; what is in need of attention is that correct data may not be accurate. For example, 011 represents for Professor, and 012 represents for Associate Professor
Staff in the standard code for appointed positions if the corresponding dictionary table; if an associate professor is represented by 011, it will not be accurate, but code 011 itself is correct.

Integrity: in a certain range, according to the specific requirements and related data, it should be ensured that the data is either not the lack of proper part or the extra incensement. For example, the data in the public database should exclude the existence of illegal values and minimize the appearance of non-null values. The former example of academic information should be undergraduate, rather than the University and the latter will bring about the emergence inconvenience for statistics and decision support.

Timeliness: represents for the timeliness of data in a certain period of time, and different data have different validity time. For example, if a system is in a long run, the database will produce an abandoned table, leaving redundancy and invalid data. At this point, these tables should be continuously monitored through changes in the amount of data to troubleshoot, excluding variations of small changes in dictionary table and volume table, and the remaining is the timeliness of the abandoned tables.

Consistency: it is a measurement whether or not the same is expressed in the specific rules of the data set. For example, in a database with "M" to express the man, "F" means that women, while in other tables will be "male", "female" to represent.

Integration: in traditional university information systems, early digital campus constructed aspects of personnel, academic, finance, and student management. However, this results in disperse data storage and processing, which makes the corresponding field of the database table in the system has no corresponding connections, even the primary key does not match or no key codes. This is the problem of data integration to solve later.

Domain Analysis and Filter Analysis

Domain analysis and filter analysis are the guarantee of data quality in the "dirty data", and data cleaning is a part of ETL technology in data warehouse. These two kinds of analysis have a great effect on the quality of controlling data. They are other methods of data quality analysis through database structure according to the definition of the data quality.

Domain analysis function is to analyze the database table through the column, including (1) Data type analysis, involving the standard amount, quantity, time, text enumeration; (2) Statistical analysis, the correlation analysis of the frequency, variance, percentage, etc.;(3) Format analysis, mainly related to pattern matching for character type correlation analysis Domain analysis can make the data environment fully understood, in order to effectively manage and assess the quality of data.

Filter analysis is mainly through the combination of different rules for evaluation. The basic rule of the filter has three types: the first is the equality rule (CONTAINS); the second is the inclusion rule (EQUALS); the three is the existence rule (EXISTS). In addition, the filter definition of the rule also has a range of rules, regular expression rules, frequency rules, type rules and unique rules, etc. The analysis and evaluation include the number of records, rules details, and the total number of records and the correct rate.

ETL Analysis Based on Data Warehouse

In each information system, there is a corresponding database system, and the database system is related to a concept of data warehouse. Data warehouse is based on database technology. It supports management for subject oriented, integrated, and
time varying data collections, which is a technology used to study how to organize data more efficiently and to facilitate the use of data.

Before data is loaded into a data warehouse, a series of operations ensure data quality, and these operations are called ETL, namely on data extraction (Extract), conversion (Transform), and loading (Load). Data cleaning ahead is mainly carried out automatically or manually to correct errors in the data, in order to improve the quality of data. ETL is actually the process of data flow, at the same time, ETL data integration services can provide the corresponding analytical functions, to achieve the analysis of integration rules. ETL is often used in large data warehouse and decision support system, and it also plays an important role in data integration of information systems.

Data Integration

Data quality analysis is finished, it is necessary to integrate data. Taking SOU as an example, it aims to provide a full range of information service support environment for university students, staff, personnel, leaders, and managers. With SOA, cloud computing, Internet of things, intelligent decision analysis as a representative of the new generation of information technology, it is urgent to create a new campus environment of network, digital, and service-oriented, in order to unify business plans, eliminate information islands, improve the existing system of open and expand a single application model.

Context of Data Integration for Service Platform

The first phase of SOU digital campus construction was established in 2003, and it realized a unified integrated construction of campus network portal, including service portal platform, unified authentication platform, public data platform, business system, etc. with the implementation of the point-to-point data exchange mode, which leads to a complicated data context, not easy to expand and coupled each repeated login system, resulting in a waste of resources.

Consequently, the data integration for the university's application system transformation and upgrading includes three aspects of information integration, authentication integration, and service integration. The information integration will define the digital campus information standards and norms, through the establishment of public database to build data exchange platform for achieving data sharing, data management and services. The authentication integration achieves a unified identity authentication, and achieves unified account management, especially the management of user rights. The service integration is mainly for the application of the system services, resource integration, and different roles for the user to provide a proactive, personalized one-stop service window.

Data Integration Design

Integrated System Framework Design. The original application system of SOU is running well, and in order not to affect the normal operation of the university's key business applications, so we use the management and service oriented architecture, and the application is loosely coupled. From the bottom to the up, the digital campus infrastructure is divided into IT infrastructure, shared data layer, application support layer, application service layer and service display layer. These five layers are shown below:
Figure 1. System frame architecture.

Figure 1 shows the system framework of the service platform. The whole university business processes, and business information storages will not be stored on a single server and database, but with the use of distributed hierarchical processing and information structure. Each application system has its own business data (resource) library and business processing, through data integration platform to achieve information sharing and exchange. Cross application sharing interoperability can guarantee the information sharing, and also in line with the business system processing distribution and information distribution.

Establish Data Integration Platform. In order to integrate application systems, the data integration platform is indeed to be established by the unified information exchange standard and data exchange interface. Through professional ETL tools, data integration platform in the business data should be processed for cleaning, extraction, conversion, loading, information exchange and sharing. Meanwhile, the technology will also provide data quality management and monitoring functions. It is also very effective on data conversion and cleaning as well as scheduling and monitoring.

Data exchange model is involved in the integration of information, and previous point-to-point data exchange interface application makes environments complex, which is detrimental to the expansion of the whole system and data sharing. After the integration, the data exchange is used bus data exchange, and the center of the "data integration platform" ensures the sharing of information, as shown in Figure 2. The integration method of bus type greatly reduces the complexity of application integration, data exchange, data exchange, and integration in the heterogeneous data source only needs to realize the integration of “data integration platform”. This method makes it possible to realize the integration of complex application systems, and also greatly reduces the implementation and maintenance costs of the whole information system.
Figure 2. Change in the pattern of data exchange.

**Requirements of Information Integration Technology**

Overall, the component-based development framework is adopted with the designing idea of SOA framework. The SOA architecture is beneficial to the construction of the project, which can be distributed, combined and used in the distributed deployment according to combination and use of the loosely coupled coarse-grained application components. The service layer is the basis of SOA, which can be directly applied to the system, so that it can effectively control the human dependence of software agent interaction in the system. J2EE technology is used in the whole technical systems, and Browser / DataBaseServer / WebServer three layer structures are adopted to develop the application system. Data underlying base is 11g Oracle database system software with RAC to set up the database cluster. Using XML data exchange technology and services of WEB SERVICES for business system integration and integrated support serves as the main way to realize the external interface of the system. In addition, it is also the assistant method as the implementation of the system data exchange.

**Conclusion**

The importance of data quality in university information systems has been analyzed, and puts forward the concept of traditional data quality analysis and data analysis as well as domain analysis and filter analysis, then proposes ETL technology and the application to data integration.

Based on the implementation of digital campus of SOU, the paper expounds the framework and technology of university information systems, the change of data exchange mode, and the relationship among integration process of data integration platform, data quality and ETL analysis technology. It realizes three aspects of data integration in authentication integration, information integration, and service integration, basically eliminates information isolation, and realizes data sharing as well as orderly, stable and efficient management and application.
References


