Research on the Construction of Big Data Platform on Smart Campus

Lina Qin

Modern Education Technology Center, Wuhan Business University, Wuhan, China

Abstract: On the basis of analysing the existed problems of big data on the smart campus, combining with the advantage of big data technology, this paper collects, processes, stores, calculates and analyses the massive scattered data of various business systems in colleges and universities. The result is regarded as the decision-making basis of teaching, research, management and logistics service in colleges and universities, so as to provide smart services for big data platform and realize the intellectualized, refined and personalized development of educational informatization.

1 INTRODUCTION

At present, the educational informatization in our country has been developed for more than 20 years, and informatization construction on the campus of colleges and university has promoted big data to enter into the campus. Big data of smart campus is a subset of educational big data. It is all kinds of data produced by teachers and students in the process of teaching, scientific research, management and logistics service, as well as the state data of various kinds of affairs management in school. Big data has the characteristics of large volume and varieties. At the same time, it is faced with such problems as lack and disorder of data, disunity in standards, isolation and so on, but it contains a lot of information and contains a great deal of value, so giving full play to its role is an indispensable part of realizing the strategic goal of school.

Under the premises of doing not change the construction mode of the existing campus information system, at the same time, making maximum use of the existing information systems and other infrastructures, the construction of big data realizes the smart operation of campus, it provides massive data analysis that supporting for campus education and teaching innovation, and promoting the transformation of school informatization from the traditional operation management to conscious service.

2 BIG DATA

Following with cloud computing and the Internet of things, big data is a disruptive technological revolution in IT industry.

With the above technologyapplication, sensor data, mobile end data, network flow data, radio frequency ID data and other data grow from TB level to PB level or even ZB level in a geometric ratio(Qi Yao, 2013). According to estimates by IDC Technology Company, in recent years, the annual data has been maintaining the 50% growth rate. The traditional relational database model cannot deal with large amount and unstructured form data, which greatly affects the embodiment of data value.

Big data technology, as a new generation of IT technology will help people extract valuable data from massive and scattered data. Internet of things, cloud computing and other technologies as the constitute part of the "smart campus", it produces a variety of isolated dispersion, collection, processing and analysis of application requirements of large-scale data make the design concept of big data constructed rapidly in the construction of "smart campus"(Qi Yao, 2013).

McKinsey, a well-known global consulting company, first proposed the come of "big data" era. It believes that the scales of some data sets are too large to be obtained, processed, stored, calculated and analysed by conventional database tools, so the company calls such data sets as big data. At the

same time, it stressed that big data does not have to exceed the TB level.

Wikipedia defines big data as a collection of data that cannot be captured, managed or processed with conventional software tools within a certain period of time (Linfei Wu, 2013).

Big data contains five characteristics, namely, large volume, high velocity, many varieties, veracity and high value, which is the "5V" characteristics.

3 ANALYSIS OF THE EXISTING PROBLEMS OF BIG DATA ON SMART CAMPUS

The existing problems of big data on smart campus are numerous and miscellaneous, they are mainly embodied invariant data pre-processing, poor data standardization and incomplete data service.

3.1 Various Methods of Data Collection, Cleaning and Warehousing Processes

With the expansion of campus scale and the increasing complexity of business, the business system information of digital campus is basically built according to the vertical line of business, the function managed by each system is different, the business data is scattered, the basic data and public data lack of synchronization and sharing and the inconsistent core model lead to data inconsistency and form the information island. In addition, the lack of planning for the source and use of data, it is difficult to analyse and apply the association or fusion intensively, so it is unable to meet the requirements of school for data analysis and decision making.

The source of big data in colleges and universities are wide, including traditional structured data and semi-structured data such as XML, as well as unstructured data in video, audio, text and other forms (Qiwei Sun, 2014). Therefore, it is necessary to make the corresponding ETL data collection and integrate all kinds of fragmented data, cleaning the data, ensuring the data quality and updating the data mode continuously for store according to the time evolution, so as to provide data analysis to the upper layer.

3.2 Poor Standardization of Data

On the smart campus, from the hardware update to the changes of various application systems, all of them promote the continuous development of educational informatization. While leaving a large amount of valuable data, it also exposes the lack of experience in data accumulation of various application systems. From the format and field of data generation to the storage, processing and analysis of the data platform, which show characteristics of single field everywhere. Some systems do not design data dictionaries, others do not provide standardized data output interfaces, and some systems only support special database storage and so on.These problems still exist objectively in big data time.

3.3 Campus Service is not Fine Enough

A wide range of sources and short time for students to get information, which make students become more personalized in their own knowledge and value system construction. At present, in college students' management, it is difficult to fully grasp the individual characteristics of students when establishing the cognition based on sensation and perception. Although the management of "mass production" is simple, it cannot meet the personalized requirements of students.

On the smart campus, the construction of a large number of business systems and the maturity of big data technology make it possible to collect total amount of the campus data. The collection of total amount of the data enables campus administrators to understand all aspects of the school's people, finances, objects and other information, and how to use the information to provide better quality, personalized and fine service for the main users of the campus----teachers and students is the advantage in application and promotion of big data.

On the basis of the above analysis, using the relevant technology to create a big data platform, collecting, processing, calculating and analysing all kinds of data on the smart campus and then turning the results into renewable resources for scientific research, management and logistics services in colleges and universities, which is the urgent need of education informatization and smartness.

4 RESEARCH ON THE CONSTRUCTION OF BIG DATA PLATFORM ON SMART CAMOUS

The construction framework of big data platform on smart campus is shown in Figure 1.
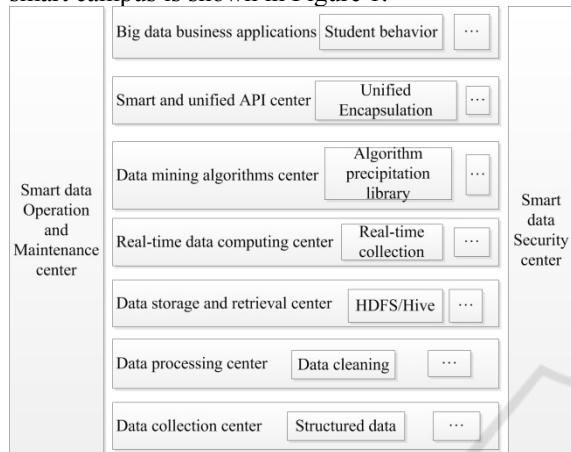

Figure 1: Construction framework of big data platform on smart campus.

4.1 Data Collection Center

Through using the collection tool flume based on log, data flow oriented collection tool Nifi, structured data oriented collection tool sqoop and ZigBee technology and other customized data collection ETL tool libraries, collecting various types of school software, hardware device data and log, Internet data and other massive scattered data, so as to support Socket, webservice, database, FTP and other common external interfaces. The data types that can be collected by the tool include distributed data information, structured data in relational databases, all kinds of semi-structured and unstructured data, static and high-low frequency knowledge data, Internet data, as well as data provided by third party partners, it can also achieve monitoring and iterative optimization of data quality.

4.2 Data Processing Center

In the aspect of data processing, the data problems often encountered are: data missing, data duplication, data error, data is not available, etc. According to different types of data problems, the following corresponding data processing methods will be adopted.

The processing methods of data missing are importing it again from the business system, manual back tracking, offsetting value according to logic, and giving up.

The processing methods of data duplication are eliminating duplication when there is full duplication, eliminating it according to time, manual removal, and eliminating it according to service logic.

The processing methods of data error are removing abnormal value by interval limit, repairing format error through planning, manual removal, and eliminating it according to service logic.

The processing methods of unavailable data are matching it according to the rule, keywords matching, enumeration conversion, and so on(Xin R, 2013).

After cleaning, processing and standardizing the existing platform data of school, the paper establishes the unified and standardized data platform, in which the standard database is set up, so as to provide the standard data access for the campus business system and formulate the detailed control strategy for access rights, so that the safety and protection of standard database is guaranteed. It realizes the standard and shared database, mainly including the organizations in the school, the basic information of teachers and students, the information of the curriculum and the building sites, as well as realizing the subscription type of data access. At the same time, it can provide a unified standard basic data platform for the construction of the post business in campus.

In the stage of data processing and standardization, at the same time, building the data standard of business system construction which is based on the national standards and school business characteristics, building the standardized management system, standardizing business description, unifying field standards, unifying data quality standards. Developing data shared format and coding standards for school information.

4.3 Data Storage and Retrieval Center

The center mainly includes the two contents of construction of data warehouse and data retrieval.

In the construction of big data warehouse, the distributed storage mode of Hadoop is mainly adopted, and three kinds of distributed storage technology are adopted to store the data in a classified way in big data warehouse platform, they are Hive, HBase and HDFS, so as to ensure the performance and requirement of the platform. For

example, for static knowledge data, that is, the real-time computing requirements are not high, mainly used to calculate the trend of data and predict data, generally for the basic data in school and storage and analysis of historical data use the Hive storage, at the same time, it supports the standard SQL language query.

The establishment processes of data warehouse are: extracting full data and continuous incremental data from the existing business system, and then the original full data warehouse is established through the storage of Hadoop big data warehouse, next, storing them into standardized database through the standardization of original data. And then the application theme database is established by modelling analysis. Then, the data in theme database is synchronized to the application access library, which provides data access for the front-end application(Scheidegger L, 2012).

The retrieval center supports the management and retrieval of the whole data warehouse.

4.4 Real-Time Data Computing Center

In the data sources of big data, many systems need to undertake real-time data collection, analysis and calculation, so as to make analysis results according to real-time data. The common data that needs real-time calculation are monitoring data, consumption data, location data, log data, and so on(Shurong Zheng, 2014). In this paper, we design a studying and processing center for real-time data, which mainly undertakes the data collection through the real-time collection tools such as Flume, scheduling bus through Kafka in real time, analysing, storing and aggregation analysis on the implementation data and other various operational processes. So as to realize the statistics of time window and online data mining as well as analysis, and finally make real-time judgment, alarm and recommendations for the system.

4.5 Data Mining Algorithms Center

After designing the data model, the business concepts, variables, and business rules have been determined, but the suitable algorithm is still needed to be chosen. This center of mining algorithm includes the algorithm precipitation library and the application model library especially for educational big data, aiming at the big data analysis system, adopting the algorithms, such as machine learning, association analysis, cluster analysis, outlier analysis and other algorithms based on the basic model and

application model, so as to realize modelling analysis for data.

4.6 Smart and Unified API Center

This center provides the unified and standard interface for data store, calls, access, and application development aiming at the big data platform. The developers can extend and develop the platform by using the corresponding interfaces.

The center supports users to access data storage platform in multiple languages, such as R language, Python, Java, SQL, and so on.

This center has the entry of application developer identity. All of the people, teams and organizations can apply independently. The administrator verifies the identity of the applicants and then automatically sends emails to inform them of their initial account number and password. Developers can manage their applications that developed by themselves, including creating applications, API applications, applicant user management, on-line applications and so on.

4.7 Smart Data Operation and Maintenance Center

Through the unified management and control on the whole data collection, data storage, data standardization, process control, automated installation, deployment and cluster of the platform, application service, security and authority in this center, it can greatly improve the efficiency of school administrators and reduce the difficulty and workload of daily operation and maintenance.

4.8 Smart Data Security Center

This platform introduces Kerberos authentication mechanism to control the grant of roles and permission, combines with multiple copies of data, data encryption technology, encryption transmission technology to ensure the secure access of the platform and the reliable guarantee of data. Furthermore, it establishes a standardized secure access system.

4.9 Business Applications of Big Data

In the application layer of big data business, the front end of this platform uses Jquery+EasyUI+Echarts components, using a large number of visual display technology, so as to show the intuitive effect of big data analysis through such as line chart, bar chart, dash board and so on, except

that, the different visual application services can be opened according to user rights.

The application of big data on smart campus is mainly reflected in teaching, scientific research, management, student behaviour, logistics service, finance, personnel, security warning and so on.

5 BENEFIT ANALYSIS ON THE CONSTRUCTION OF BIG DATA PLATFORM

This paper analyses benefits on the construction of big data platform from economic aspect and social aspect.

Standing on the economic slide, the business system, hardware resources and other data can be collected on the big data platform, and users can used these data through Internet access only according to their own application requirements. In addition, it can provide data retrieval quickly, through the data warehouse to save the stock data and incremental data, at the same time, achieving the accumulation and backup of the data. On campus, the big data collects and analyses the educational data produced in the educational activities, and the results are taken as the basis of the decision making in Department of Management and the investment of manpower and material resources are reasonably reduced.

Social benefits are shown in the following education aspect, security aspect, and service aspect.

In the aspect of educational decision - making, whether in helping the decision - makers to understand the present situation more clearly, grasping more comprehensive and more valuable information, or in the process of formulating, implementing, and adjusting the specific educational policies, the big data of education still plays an important role in these aspects.

Big data platform can realize the goal of comprehensive prediction of campus safety by forming a comprehensive and dynamic student behaviour monitoring system, analysing and predicting students with abnormal behaviours, so as to notify the school authorities in time to avoid accidents in advance, and maintaining the image of school.

Big data technology can improve smartness of educational management and decision-making, and become an important power to promote the development of smart education.

6 CONCLUSIONS

In the paper, the big data platform effectively collects and integrates the data of all kinds of business systems and hardware equipment in colleges and universities, solves the disadvantage of traditional relational database storage and dealing with many kinds of morphological data, provides a unified interface for secondary development and application, provides smart service for teachers, students and administrators in the school, helps to achieve the intellectualized, refined and personalized development of smart education.

REFERENCES

1. Qi Yao, 2013. Research on the value of big data on the smart campus.*Journal of Information-Network Security,* 152(8) August, pp.91-93.
2. Linfei Wu, 2013. Management of radio and television network customer relationship in big data era.*Journal of Radio and Television Information,* 258(10) October, pp.36-39.
3. Qiwei Sun, Chun Lu, 2014. Research on the application of big data in colleges and universities.*Journal of Chinese Educational Network,* (1) January, pp. 63-65.
4. Xin R, Gonalez J, Franklin M. 2013. Graph X: A resilient distributed graph system on spark.*Proceedings of 1st International Workshop on Graph Data Management Experience and System*, New York, International Financial, pp. 12-18.
5. Scheidegger L, Vo H T, Kruger J, et al. 2012. Parallel large data visualization with display walls.*Proceedings of the 2012 Conference on Visualization and Data Analysis*, USA, Lancaster, pp. 1-8.

6. Shurong Zheng. 2014. Formation, application and enlightenment of big data in retail business. *Journal of Theoretical Exploration,* 206(2) February, pp. 90-94.