A Campus Big-Data Platform Architecture for Data Mining and Business Intelligence in Education Institutes

Ningcheng ZHANG

Shanghai Jian Qiao University, Shanghai, China

zhangningcheng8@sina.com

Keywords: Campus Data, Business Intelligence, Big Data, Data Mining.

Abstract. Although the commercial business area has witnessed successful applications of BI technologies, in the noncommercial education institute, BI is still in its early adoption phase, partly due to the budget constraint. In addition, the introduction of new Internet-of-Things technologies in campuses brings unprecedented rich data that reflect students' activities in their campus life. The big data provides valuable insight for further improvement of the campus management as well as strategic decision-making of the institutes. On the other hand, it also sets forward challenges on the traditional campus information system. In this paper, we present our big data platform architecture design for the BI application system on the basis of the successful project experience of the university. The architecture has the outstanding advantages in the cost, expandability and scalability. The realized BI analytical functions have played a significant role in the decision making in the university's strategic and routine management.

Introduction

Business Intelligence for Education Institutes

The explosive developments of the science and technologies are making in-depth impacts on today's industry and society. More and more industry sectors are experiencing significant transformation and upgrade. Many traditional working posts become out-of-date, while new talent requirements are opened in burst. At the same time, the new generation of youngster has more diverse and pronounced individuality and character. They have generally high expectations on the resources and quality of education institutes. All these put forward new challenges for education institutes in providing up-to-date skill training and improving the teaching/training methodologies that are adaptive to students and trainees.

Taking advantages of the latest data management and analysis techniques, business intelligence (BI) system is able to explore strategic and operational business knowledge and insight, e.g., the market prediction or customer profiles/preferences, by integrating and mining a wide spectrum of enterprise data together with other relevant information [1-2]. Although the commercial business area has witnessed successful applications of BI technologies, in the noncommercial education institute, BI is still in its early adoption phase [8], partly due to the budget constraint. Nevertheless, it is widely recognized that the BI and its related data mining technologies will play an essential role in help modern education institutes tackle the above challenges in this fast-changing world.

Internet-of-Things Application in Campus

In recent years, Internet-of-Things (IoT) technologies become more and more popular in almost every industry domains. By embedding sensors into front field environments as well as terminal devices, IoT network is able to collect rich sensor data that reflect the real-time environment conditions of the front field and the events/activities that are going on. Advanced data mining technologies can be applied to explore in-depth business insights from these data. Since the data is collected in the granularity of elementary event level in a 7X24 mode, the data volume is very high and the data access pattern also differs considerably from traditional business data. This has motivated a new generation of data management solution, e.g., NoSql database, map-reduce distributed computing framework, etc.

IoT technologies have already a lot of applications in the universities in China. In many

universities, for example, the so-called campus card system has been more and more popular nowadays. The campus card is like a unique identity card of a student or a university employee. The card is uniformly used in accessing almost all university facilities like dormitory, library, laboratory, dining hall, restaurant, and etc. Whenever a card is used, the terminal card read device generates an event message and sends it to the back-end server of the campus card system, which in turn records the corresponding data in the back-end database. Such kind of data actually reflect daily activities of the card users, typically the students. With advanced data mining technologies [7, 9-20], value insights can be explored for campus management as well as strategic decision-making for the universities administration board. Correspondingly, the campus information sys-tem should be able to accommodate the management and usage of such data.

Structure of the Paper

In this paper, we propose a campus big-data platform architecture supporting advanced business intelligence in typical education institutes, on the basis of the project experience in the information system upgrade in the Shanghai Jian-Qiao University. Education institute in this paper refers mainly to, but not necessarily limited to the universities or colleges.

The rest of the paper is organized as follows. We will present the requirement analysis of the campus information system in the BI realization. Then the system architecture is introduced and explained in details. Example BI analytical functions based on this architecture are presented. At last, the work is summa-rized.

Requirement Analysis

Aiming at establishing an information system for BI applications, the following fundamental requirements are discovered with respect to the status of the legacy campus information system.

Data Conversion and Integration

Since most campus information systems were de-veloped in an incremental way, today's overall campus information system is generally segmented into several subsystems, typically the student management subsystem, the human resources (HR) subsystem, the library management subsystem, etc. Each subsystem corresponds to a specific campus management system and was developed and is operated independently. For an effective data analyses in BI modules, the data from these different subsystems need to be integrated and, if necessary, transformed for the consideration of consistent syntax and semantics.

Big Data Storage/Processing

In China, most university/college campuses are some kind of micro-cities that incorporate also living and entertainment facilities in addition to the education resources and facilities. These facilities are uniformly under the management of the university or college. To simplify the access and improve the utility, more and more universities and colleges introduce the campus card system. A campus card is a unique card that identifies a student or an employee. A student or an employee needs to use his/her own campus card to access the campus facilities (e.g., dormitory, library, laboratory, gymnasium, etc.) as well as for any payment for consumption (e.g., dinner, shopping, entertainment, etc.) in the campus. The card system records every usage of the campus card and distributes the records to different subsystems like identity authentication subsystem and payment subsystem. Besides, the card system also keeps a copy of the overall card usage data for audit purpose. In a middle-sized university or college with more than ten thousand stu-dents/employees, the daily campus card usage data can amount to hundreds of thousand pieces of records. Growing with the time, that is a considerable data amount. The current relational database (RDB) system is getting cumbersome in handling such kind of data, in consideration of the system scalability and data access pattern of these data (i.e., frequent write and relatively rare query) [3]. Correspondingly, the processing of such large amount of data requires also economic and scalable distributed computing architecture.

Flexible Multi-Dimensional Analysis Data Model

In order to flexibly support a wide range of analysis scenarios as well as a good response performance in generating and visualizing the analytical results, raw data from campus information

subsystems and campus card system need to be processed, extracted and ag-gregated to an appropriate abstraction level. Then the data need to be organized into data model specially tailored for multi-dimensional data analysis.

Proposed Data Platform Architecture

Following the above requirements, the campus data platform architecture is proposed, as illustrated in Fig.1.

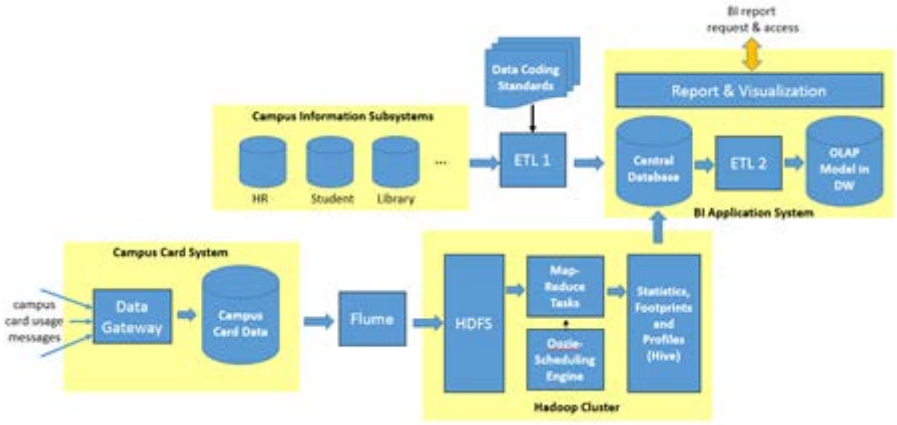

Fig.1: Proposed Data Platform Architecture

The left part of Fig.1 includes the legacy campus information subsystems and the campus card system. Each of these subsystem is independently administrated and operated for specific functionalities.

To build up the data analysis platform supporting business intelligence applications, two additional systems are built. A Hadoop cluster is constructed. It stores a copy of the campus card data and executes distributed map-reduce tasks to extract useful information for decision-making.

The BI application system collects processed in-formation from the campus information subsystems and the campus card system respectively. These data are further aggregated and reorganized into the multi-dimensional data models. BI report and visualization functionalities are built on top of the data model.

In the following subsections, we will focus on the introduction of the Hadoop cluster system and the BI application system as well as the data transform/import from the campus information subsystems and the campus card system.

Data Transport/Transform from Campus Information Subsystems (ETL-1)

To meet with the first requirement (data conversion and integration) in Section 2.1, the relevant data in the current individual information subsystems are extracted, transformed and loaded (ETL) into a Central Database, the process of which is performed by the component ETL-1 in Fig. 1. During the ETL process, different data schema and coding standards are applied to transform the data of heterogeneous sources to a uniform representation. They includes the State standard (e.g., the coding scheme for countries, states, nations, gender, etc.), the standard of the Ministry of Education (typical coding scheme for majors, disciplines and programs) and the university/college standard (e.g., coding scheme for departments, organization, etc.), following a descendent priority order.

Hadoop Big Data Platform

With respect to the second requirement (big data storage and processing), a distributed storage/processing platform is proposed. For that purpose, the open source Hadoop cluster is adopted [3]. Hadoop is a distributed data storage and processing platform that is designed for big-data-based applications. On this basis, a full set of open source components are available. In the following subsections, we discuss the architecture design for data storage, data processing, ETL and data flow management over the Hadoop platform.

Data Storage and Schema. For the data storage, Hadoop provides two basic approaches: Hadoop Distributed File System (HDFS) and no Sql database Hbase. HDFS is actually the fun-damental file

system of the Hadoop system. Files in the HDFS are transparently saved into several copies that are distributed to different server nodes. In this way, HDFS provides high availability and reliability of the data.

Hbase is one of the most popular NoSql database. It uses column-based data storage structure, which makes it very easy to add additional data fields in the future without restructuring existing data schema. A very promising feature for the system expandability.

A decision on the data storage between HDFS and Hbase needs to consider the following pros and cons of the both approaches. Comparatively, HDFS is a low level storage approach and thus has the advantage of fast writing/reading speed when the data are saved/read in a batch of files. However, HDFS does not provide data query mechanism. On the contrary, Hbase provides better support in data query according to table keys. However, the data reading/writing speed is not comparable to that of the HDFS. In our scenario, the data injection speed of the campus card data is not extremely high and there is indeed requirements on the query of the raw campus card data. For this reason, we choose Hbase as the data storage approach in the Hadoop big data platform.

In Table 1, a set of typical data fields of the Hbase data schema for the campus card data is presented. The key of the table is a combination of the unique card ID and the card usage timestamp.

Table. 1 Hbase data schema for campus card data

Field Name	Type	Description
Card ID + Transaction Timestamp	String	Table Key
XGH	String	User ID
JYDD	String	Transaction Location
JYLX	String	Transaction Type
SHMC	String	Transaction Entity
POSJH	Number	Transaction Device ID
JYJE	Number	Cost Amount
KPYE	Number	Card Residual Amount

Note in Table 1 that the names of individual fields are relatively short. This is because that Hbase tables are physically organized in terms of key value pairs. Shorter field names lead to shorter key length and can bring significant storage saving.

Footprint Analysis with Map-Reduce Tasks. It can be seen from the data schema that the raw campus data set is a collection of trivial log data that do not disclosure much useful information directly. Advanced data processing is necessary for that pur-pose.

Within the Hadoop cluster, a series of map-reduced tasks are realized, which aims at analyzing the raw card data to generate enriched information that is able to provide business insight. Typically, it includes the footprints, statistics and profile information, which are introduced as follows.

Footprints -- The footprints here mainly refer to a trajectory of facilities which a student daily visited. For example, a typical student's timestamped footprint trajectory can be classroom (7:55) – canteen (11:30) – dormitory (12:15) – gym (15:30). This can be directly derived from the facility access information of the card data. With the footprint trajectory, the staying time at each facility can be approximately estimated by the time interval between two adjacent visits. Besides, in case that the visit of a facility is associated with an expense (e.g., shop, canteen, etc.), the consumption information is also recorded accordingly (cf. Table 1).

Statistics – Based on the raw card usage data and the footprint information, different statistics of a student can be calculated. Typically, this includes the visit times and sojourn duration at different kinds of facilities, the total expense and percentage of the expense for different purpose, etc.

Profiles – Based on the above information, it is possible to segment students to different classifications on different dimensions. For example, a high frequency of visiting the library indicates a reading-lover, while a regular visitor of the gymnasium indicates a gym-lover. Besides, frequent shopping records late at night can be a sign of a nightlife lover, while frequent laboratory or study room access at midnight is a clear symbol of a night worker.

Analytical Result in Hive. The analytical results from the above map reduce tasks are stored in the

HDFS file. To provide a flexible query capability of the results, it is desirable to organize the results in a more structured way.

In our implementation, Hive [3] is adopted for that purpose. Hive provides the data modeling and pro-cessing framework for HDFS data and supports SQL-like data query, which fits our architectural re-quirement very well.

Processing Workflow Scheduling. The map reduce tasks in Section 2.2.2 have different purposes and are supposed to be executed over different time periods. Footprint analysis, for example, can be executed on a daily basis. Statistics and profiles are typically checked on the monthly basis.

The execution of the map-reduce tasks is scheduled by the Oozie scheduling engine [5]. Two time-based Oozie flows are defined. One is activated on a daily basis and the other is on monthly basis, realizing the above task scheduling requirement.

Campus Data Import via Flume. The card data is routed to the Hadoop cluster from the campus card system via the data import middleware of Flume [4]. In the Flume component, data format verification and cleaning are implemented to screen out those error or invalid card usage messages. Cleaned card data are written to the HDFS storage upon entering the Hadoop cluster.

Besides the support for the above data preprocessing, Flume has also the advantage of the support on importing event-based messages, which makes it also suitable to serve as a message middleware for near real-time data applications. In addition, Flume can be easily scaled out when the data volume and throughput increase in the future. All these make Flume a good choice for the data platform's expandability.

BI Application System

The overall results of the Hadoop data analytics are exported by tool like Sqoop [6] to the Central Database. The Central Database is a relational database located in the BI application system. It collects all the necessary data from the subsystems. The BI report and visualization module can directly fetch the data from this database. Some data still needs to be processed and organized into more appropriate form for more flexible data query and analysis.

A typical model for this purpose is the OLAP (online analytical processing) model [7]. This model organizes interested measures or metrics (called facts in OLAP terminology) with reference to multiple analytical dimensions, so as to avail flexible analytical views of the metrics over individual dimensions or an arbitrary combination of individual dimensions. In Fig. 2, an OLAP data model for the course score is sketched. The CourseScore table here is the fact table containing two interested metrics: the number of the student's absence in this course and the final exam score. Through the four foreign keys, studentID, courseID, lecturerID and semesterID, this fact table is associated to the four dimension tables: Student, Course, Lecturer and Semester, which is a typical star schema [7] of OLAP model.

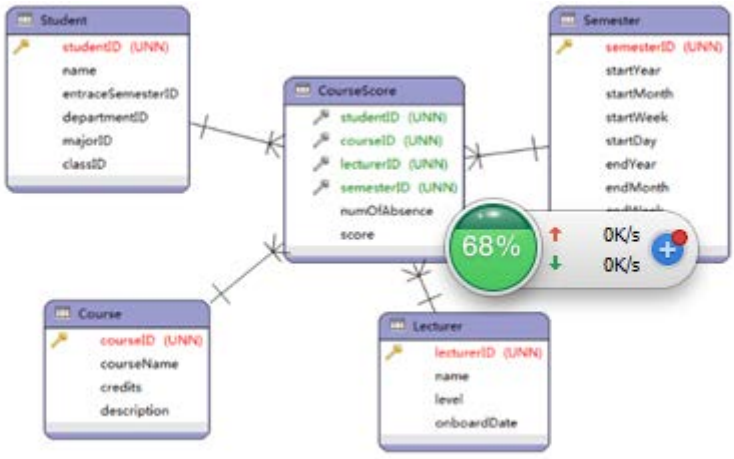

Fig.2:Star schema of course-score data model

The processing component ETL-2 conducts the OLAP processing and stores the resulting data model into the data warehouse (DW), which is also a relation database. The DW is another direct

data source for the report and visualization module.

Analytical Function Examples

In this section, two analytical function examples of the BI application system are presented, as illustrated in Fig. 3 and Fig. 4.

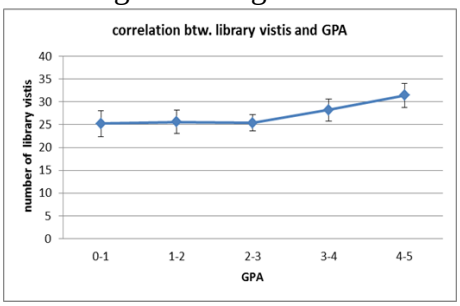

Fig. 3: Correlation btw. GPA and library visits

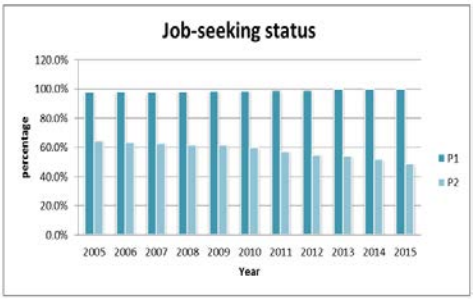

Fig. 4: Tendency in job-seeking status

Fig. 3 shows the correlation of students' GPA (grade point average) with the number of the times of library visits. It can be seen that students with GPA above the average do have a higher frequency in the library visits.

Fig. 4 shows the job-seeking status when the gradu-ates of a specific college major left the university. Two metrics are designed for the job-seeking status: the percentage of the graduates that got a job offer before they left the university (denoted by P1), the percentage of the graduates that got a job position matching well with the major (denoted by P2). It can be seen that since 2008 the percentage P1 remains at the same level while P2 shows an obvious decrease. This reflects the tendency of the talent requirement for this major on the human resource market, which is an important signal for the necessity of tuning this major's enrollment plan and course program.

Summary

Although the commercial business area has wit-nessed successful applications of BI technologies, in the noncommercial education institute, BI is still in its early adoption phase, partly due to the budget constraint. In addition, the introduction of new Internet-of-Things technologies in campuses brings unprecedented rich data that reflect students' activities in their campus life. The big data provides valuable insight for further improvement of the campus management as well as strategic decision-making of the institutes. On the other hand, it also sets forward challenges on the traditional campus information system.

In this paper, we present our big data platform ar-chitecture design for the BI application system on the basis of the successful project experience of the uni-versity. The architecture has the outstanding advantages in the cost, expandability and scalability. The realized BI analytical functions have played a significant role in the decision making in the university's strategic and routine management.

Acknowledgement

This work was financially supported by the Research Foundation of Shanghai Jian Qiao University.

References

[1] D. Arnott and G. Pervan: "Eight key issues for the decision support systems discipline", Journal of Decision Support Systems, Vol. 44, No. 3, 2008, pp. 657–672.

[2] J. Luftman and T. Ben-Zvi: "Key Issues for IT Executives 2009: Difficult Economy's Impact on IT", MIS Quarterly Executive, Vol. 9, No. 1, pp. 49–59.

[3] Tom White, Hadoop -- The Definitive Guide, O'REILLY, 2012.

[4] Apache Flume, https://flume.apache.org/

[5] Apache Oozie, http://oozie.apache.org/

[6] Apache Sqoop, http://sqoop.apache.org/

[7] Jiawei Han, Micheline Kamber, Jian Pei, Data Mining -- Concepts and Techniques, Morgan Kaufmann, 2012.

[8] Adhi Nugroho Chandra and Yohannes Kurniawan: “A Model of Business Intelligence Systems for School: A Case Study”, Springer Lecture Notes in Electrical Engineering, Vol. 331, 2015

[9] Preeti Mulay and Parag A. Kulkarni: “Knowledge augmentation via incremental clustering: new technology for effective knowledge management”, Journal of Business Information Systems, Vol. 12, Issue 1, 2015

[10] Murtadha M. Hamad and Banaz Anwer Qader: “Knowledge-Driven Decision Support System based on Knowledge Warehouse and Data Mining for Market Management”, Global Journal of Management and Business, Vol. 13, No 10-E, 2013

[11] Adrian Solomon and Panayiotis Ketikidis and Alok Choudhary: “A knowledge based approach for handling supply chain risk management”, Proceedings of the Fifth Balkan Conference in Informatics, pp. 70-75, 2012

[12] Feng Chen and Pan Deng and Jiafu Wan and Daqiang Zhang and Athanasios V. Vasilakos and Xiaohui Rong: “Data mining for the internet of things: literature review and challenges”, International Journal of Distributed Sensor Networks, 2015

[13] Behzad Beheshti and Michel Desmarais, “Predictive performance of prevailing approaches to skills assessment techniques: Insights from real vs. synthetic data sets”, Proceedings of the 7th International Conference on Educational Data Mining, pp. 409-410, 2014

[14] Cristobal Romero and Sebastian Ventura, “Data mining in education”, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 3, No. 1, pp. 12-27, 2013

[15] Katrina Sin1 and Loganathan Muthu: “Application of data mining in education data mining and learning analytics – a literature review”, ICTACT Journal on Soft Computing: Speical Issue on Soft Computing Models for Big Data, Vol. 5, Issue 4, 2015

[16] Abeer Badr El Din Ahmed and Ibrahim Sayed Elaraby,“Data Mining: A prediction for Student's Performance Using Classification Method”, World Journal of Computer Application and Technology, Vol. 2, No. 2, pp. 43-47,2014

[17] Ryan Shaun Baker and Paul Salvador Inven-tado,“Educational data mining and learning analytics”, Learning Analytics, pp. 61-75, 2014

[18] George Siemens and Ryan S. J. D. Baker, “Learning analytics and educational data mining: towards communication and collaboration”, Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, pp. 252-254, 2012

[19] Siti Khadijah Mohamad and Zaidatun Tasir, “Educa-tional data mining: A review”, Procedia-Social and Behavioral Sciences, Vol. 97, pp. 320-324, 2013

[20] Alejandro Pena-Ayala, “Educational data mining: A survey and a data mining-based analysis of recent works”, Expert systems with applications, Vol. 41, No. 4, pp. 1432-1462, 2014