Big Log Analysis for e-Learning Ecosystem

Qinghua Zheng, Huan He, Tian Ma, Ni Xue, Bing Li, Bo Dong
SPKLSTN Lab, Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China
qhzheng@mail.xjtu.edu.cn; hehuan@mail.xjtu.edu.cn; sky.ma@stu.xjtu.edu.cn;
xn_shelly@qq.com; cslibing@gmail.com; dong.bo@mail.xjtu.edu.cn

***Abstract*—Recently, e-Learning emerges a rapid development, and e-Learning ecosystem has been proposed to provide sustained and stable services to cope with the growing demands of e-Learning. In e-Learning ecosystem, the amount of e-Learning log grows exponentially, introducing the variety and complexity of e-Learning log analysis. Therefore, a robust, scalable and practical logging architecture is urgently needed. Firstly, the characteristics of e-Learning log and its analysis are studied in this paper. Specifically, e-Learning log implies complicated characteristics, such as multi-dimensional correlation, heterogeneous multi-source, and cascading generation. Furthermore, the log analysis represents abundant diversity of demands, variety of methods and low-latency requirement in computation. Thereupon, this study presents a comprehensive logging architecture covering the whole life cycle of e-Learning log data which includes log collection, transport, storage, computation and service. To verify the proposed logging architecture, a related experimental implementation is developed for a realistic e-Learning ecosystem, and three typical e-Learning analyses are proposed.**

Keywords—e-Learning ecosystem; e-Learning log data; logging architecture; characteristics of e-Learning log data and log analysis

I. INTRODUCTION

Along with the evolution of cloud computing and mobile computing, e-Learning emerges a rapid development [1, 2] recently. Online learners around the world can achieve "anytime, anywhere, any device" learning to satisfy their individual demands [3, 4]. In order to cope with the growing demands for e-Learning, an e-Learning ecosystem was proposed to provide sustained and stable service for learners, and complete technical supports covering all layers and entire learning process [5, 6].

In an e-Learning ecosystem, the log data is the lifeblood of services for learning and teaching activities based on the log analysis. However, there are three challenges to take full advantage of e-Learning log data: (1) It is difficult to model the log data due to its multi-dimensions including behavior, time, space, knowledge map [7], learning resources and learner groups etc. (2) The storage system is strictly required for processing massive log data with various sources, formats and applications of the ecosystem. (3) Considering the complexity and variety of the demands for log analysis, effective analysis methods and tools for log data are urgently required. Therefore, the log analysis of an e-Learning ecosystem is regarded as the "big log analysis"

In this paper, a logging architecture is constructed for an e-Learning ecosystem to cope with the challenges mentioned above. Firstly, the characteristics of e-Learning log data and log analysis for various users are studied. Secondly, a logging architecture is proposed according to the characteristics studied above. Thereupon, the proposed logging architecture is applied for a realistic e-Learning ecosystem, i.e. BlueSky e-Learning system [8], and three typical analysis scenarios are depicted to verify the effectiveness of the logging architecture.

The major contributions of this paper are summarized below.

1. A logging architecture which covers the whole life cycle of log data including log collection, transport, storage, computation and service is proposed to satisfy various demands for log analysis according to the characteristics of log data and log analysis.

2. Three typical scenarios are demonstrated based on the proposed logging architecture implemented on the BlueSky e-Learning system. In addition, the procedures of the batch processing, stream computing and iterative computing for log data are illustrated respectively.

The rest of this paper is organized as follow. Section II discusses related work. Section III analyzes the characteristics of log data and log analysis in an e-Learning ecosystem. Section IV specifies the proposed logging architecture. Sections V demonstrates three typical scenarios of big log analysis on a real e-Learning platform. Section VI concludes the paper.

II. RELATED WORK

Along with the rapid development of e-Learning, the research on e-Learning data expands increasingly. In terms of log analysis, educational data mining is mainly focused on[9,10]. Jong described a learning diagnostic system that collects learning logs of students and determines abnormal learning status of students to send warning messages [11]. By employing usage analysis, some rules of behaviors of course participants could be found from the questionnaires and log files of e-course [12]. Mahajan introduced an adaptive e-Learning environment to help teachers evaluating learning process with log data [13]. The effects of different teaching arrangements on learning process could be recognized by analyzing the operating records of students while they are watching video [14]. Klasnja-Milicevic et al. described a recommender system which recognizes different patterns of learning style and learners' habits by mining their server logs [15].

The log data used in research mentioned above is mainly from the application layer according to the architecture of an e-Learning ecosystem. However, along with the development of mobile learning, not only in content layer and infrastructure layer, but also in mobile devices produces important data.

Therefore, how to build an efficient logging architecture for big log data of an e-Learning ecosystem is urgently required to obtain abundant valuable information.

In the internet industry, along with the widespread application of big data, a lot of enterprises built a logging and analytic system within their business ecosystem which aggregates and analyzes log data and business data to provide support for business or using the analysis result as a new business. Thusoo described the data warehouse and analytics infrastructure at Facebook, they use Scribe server aggregates the logs coming from different web servers and load data from MySQL to the Hive Hadoop clusters, which is also used for querying and analyzing data predominantly. The results of data analysis could be loaded back to MySQL in order to serve Facebook users [16]. Abraham, et al introduced a new analysis framework at Facebook: Scuba, which was built and used for most real-time, ad-hoc analysis of arbitrary data. The most common use of Scuba is for real-time performance monitoring, trend spotting, product analysis and pattern mining [17]. Similarly, a unified logging framework and an analyzing framework were built at Twitter to provide batch and real-time data processing [18, 19, 20]. At LinkedIn, there are also some experiences in making use of user data and log data to provide references information to users with analyzed data [21].

As described above, some comprehensive architecture for log data analysis in industry have been achieved, however, the characteristics of log data and log analysis in e-Learning ecosystem are different. Thus, in the next section, these characteristics are studied to design a suitable logging architecture.

III. CHARACTERISTICS OF BIG LOG ANALYSIS

The log data and the demands of log analysis for an e-Learning ecosystem, compared with those for other application systems, represent distinctive characteristics which are illustrated as follow.

A. Characteristics of log data

Firstly, e-Learning ecosystem log data focuses on learners and contains several dimensions including behavior, time, space, knowledge map, learning resource, and group and so on, represents multi-dimensional correlation. Each dimension of log data performs its own rules, and is not only correlative but also influenced and restricted with other dimensions: 1) Different behaviors of learners could show continuity in time and space, connectivity on knowledge map, locality when achieving access to learning resources, and transmissibility in learner group; 2) Each record of log data is directly related to time, and knowledge map correlates with learning resource; 3) Different behaviors of learners produce different log data; 4) During a certain learning process, the learning behavior of learners and the accessed learning resource would be constrained to knowledge map and learner groups where learners are located at. Therefore, it is obvious that the dimensions of log data contain abundant learning pattern information.

Secondly, e-Learning ecosystem log data possesses heterogeneous multi-source. Considering the development of mobile learning and the hierarchic architecture in [6], in this paper, the data source of log data for an e-Learning ecosystem is divided into four layers: Infrastructure Layer, Content Layer, Application Layer, and Terminal Layer, which cover the overall ecosystem from cloud to terminal. In addition, since the log data performs differently in content structure and storage format, it is necessary to preprocess log data into normalized form before analyzing it to reduce processing time and save resources. The sources of log data and possible objects in each layer are listed in Table I.

TABLE I. SOURCES OF LOG DATA IN E-LEARNING ECOSYSTEM

Layers	Sources	Objects
Infrastructure Layer	Power cost, system load, CPU utility, memory usage, hard disk I/O, network traffic, etc	Server hardwares, firewalls, virtual machines, etc
Content Layer	Web request log, video usage log, learning materials, learner information, etc	HTTP servers, Streming Video Server, CDN servers, database, etc
Application Layer	Operation log, learner assessment, examination content, forum discussion, experiment results, etc	LMS(Learning Management system), content management system, forum, virtual laboratory, etc
Terminal Layer	Learner operations, environment preferences, web page or application running performance, etc	Desktop client software, web browser, mobile phone, tablet, etc

Thirdly, e-Learning ecosystem log data is cascading generated. As for multi-dimension and multi-source of log data, a certain behavior of learners at Terminal Layer could be transferred layer-by-layer through Application Layer and Content Layer to Infrastructure Layer eventually, which stimulates responses of all objects in each layer to generate corresponding log data. The amount of log data shows an exponentially growing with the increase of the number of learners and their learning behaviors.

According to the characteristics of log data in an e-Learning ecosystem analyzed above, it is prominent that the log data reveals the characteristics of big data. Therefore, it is reasonable and feasible to apply the analysis methods and tools of big data for log data.

B. Characteristics of log analysis

On the basis of the characteristics of log data analyzed above and various demands towards log analysis of different users, the characteristics of log analysis are studied as follow.

Firstly, the demands of log analysis represent abundant diversity. The log analysis demands of different roles people played in each layer are distinctive. Furthermore, log data depicts the complete status of an e-Learning ecosystem, thus each role can obtain valuable information they require through analyzing log data. The demands of different roles for log analysis in each layer are listed in Table II.

Secondly, there exist various log analysis methods towards different computation framework. In order to satisfy

TABLE II. DEMANDS FOR LOG ANALYSIS IN E-LEARNING ECOSYSTEM

Layers	Roles	Demands
Infrastructure Layer	Data center administrator	System load prediction, VM scheduling, power consumption optimizing of computing center, network bandwidth deploy
Content Layer	Server administrator	Learning materials scheduling, web cache optimization, web server and database performance optimization
	Developer	Web page structure optimization, improving GUI design, software bug detection
Application Layer	Teacher	Evaluation of learning performance and teaching performance, online response statistics
	Staff	Student enrollment statistics, curriculum statistics
	Researcher	study learning style, learning pattern, etc
Terminal Layer	Learner	Course guidance, personalized recommendation

different demands, it is necessary to construct appropriate computation framework which composes of several sufficient log analysis methods with specified metrics to evaluate collected log data for significant information. For instance, when establishing a computation framework of an learning evaluation model for learning effects of learners, not only a statistic analysis method to estimate the overall learning effects for an entire learning group, but also a data mining method to analyze the learning history of each learner are required. Moreover, taking the characteristics of courses into consideration, a comprehensive evaluation model of learning effects is constructed based on the computation framework mentioned above.

Thirdly, the process and application of log analysis require low-latency which is measured by computation time of implementing the computation framework. Whereas, the lowest acceptable computation time for different users are not the same. Specifically, for researchers, a few minutes' computation time or even a few hours' are acceptable. Comparatively, for the administrator of Data Center who needs to know about the current statistic status and future variation trend of a system in time, the computation time within one minute is acceptable. Moreover, the teachers who attend a live classroom require obtaining the latest responses of students about their learning effects in a few seconds.

Thereupon, a comprehensive e-Learning ecosystem-oriented logging architecture is constructed on the basis of the characteristics of log data and log analysis as follow.

IV. LOGGING ARCHITECTURE FOR E-LEARNING ECOSYSTEM

According to the characteristics analyzed above, a logging architecture which covers the whole life cycle of log data is designed, shown in Fig. 1. The logging architecture consists of five modules: Collection Module, Transport Module, Storage Module, Computation Module and Service Module. The function of each module is described as follow.

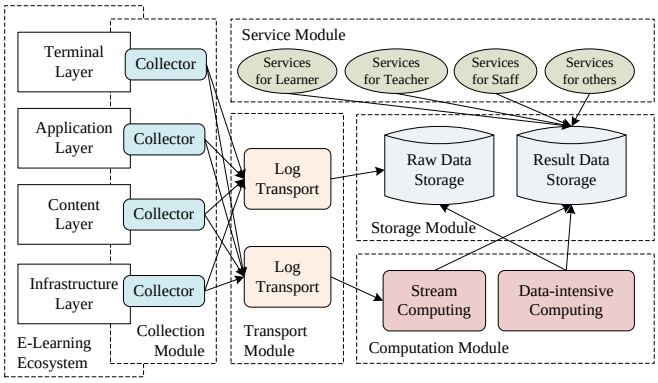

Figure. 1. Logging architecutre for e-Learning ecosystem

- The Collection Module contains collectors distributed in each layer. As log data possesses heterogeneous multi-source, each collector records the log data produced by different objects and normalizes the collected data which is shown in Table 1.
- The Transport Module transfers collected log data to where the data is required. As for the log data is usually temporarily stored separately when being collected, it could bring great cost if processing collected log data immediately, which is fairly disadvantageous for analyzing big log data effectively. Therefore, it is necessary to transfer log data to the Storage Module or the Computation Module for centralized processing.
- The Storage Module includes two categories of storage systems. One is Raw Data Storage system, which stores historical data orderly and permanently for future data mining and analyzing. This storage system should have excellent scalability to cope with the increasing massive log data from each layer of e-Learning ecosystem, and support data-intensive computing to satisfy the demands of various roles. The other is Result Data Storage system, which has the ability of rapid access to data to provide high I/O performance for stream computing in the Computation Module and the Service Module.
- The Computation Module implements the calculation process for log analysis. In this paper, the designed Computation Module contains two computing frameworks to satisfy the basic analysis demand and real-time demand. To be specified, the basic One is data-intensive computing framework applied to analyze massive raw data to dig out valuable information. The other is stream computing framework applied to deal with every coming data in real-time, aiming at quickly receiving responds or discovering exceptions. In addition, each computing framework is required to support parallel computing to guarantee low latency of analyzing process and

TABLE III. AVAILABLE OPEN-SOURCE SOFTWARE FOR EACH MODULE

Architecture Module	Open Source Software
Collection and Transport Module	Apache Flume, Apache Kafka, Flunted, Scribe, LogStash, Rsyslog, rsync, etc
Computation Module	Data-intensive computing: Apache Hadoop MapReduce, Apache Drill, Apache Hive, Apache Mahout, Apache Spark, Apache Pig, etc Stream computing: Apache Spark Streaming, Storm, etc
Storage Module	Raw data Storage: Apache Hadoop HDFS, Apache HBase, etc Result data storage: Apache CouchDB, MongoDB, MySQL, Redis, etc
Service Module	Apache CXF, Apache Struts, etc

improve computational efficiency. Furthermore, the Computation Module is extensive for its computing frameworks based on actual analysis demands. For instance, a iterative computing framework would be extended into the Computation Module to achieve high performance computing.

- The Service Module provides service for diverse users. Each service reads needed data from the Result Data Storage system for all objects and roles in each layer of an e-Learning ecosystem. In terms of functionality, the services in Service Module can be classified into three types: historical log data statistics services, real-time monitor services and prediction services. Each demand of the roles in e-Learning ecosystem may involve some services of each type. For example, to meet the demand of an administrator in data center for virtual machine scheduling, three services are needed: a service of the historical log data statistics of learning material usage, a service of monitoring accessing in real time, and a service of predicting the trend of system load.

According to our investigation, most basic architectures can be readily constructed on the basis of open-source software which is listed in Table III. In order to verify the availability of the proposed logging architecture, it is implemented on the BlueSky e-Learning system and illustrated by several scenarios.

V. APPLICATIONS OF LOGGING ARCHITECTURE

The experimental implementation of the proposed logging architecture is based on BlueSky e-Learning ecosystem which is developed by Distant Learning College of Xi'an Jiaotong University. In addition, it is applied in three typical scenarios related to teaching activities.

Each scenario specifies log data processing with different frameworks of each module in the logging architecture, including computing students' admission and attendance situations in batch processing framework, monitoring students' learning performance on Live Classroom based on stream computing framework, applying iterative computing framework to achieve the evaluation of student attitudinal performance.

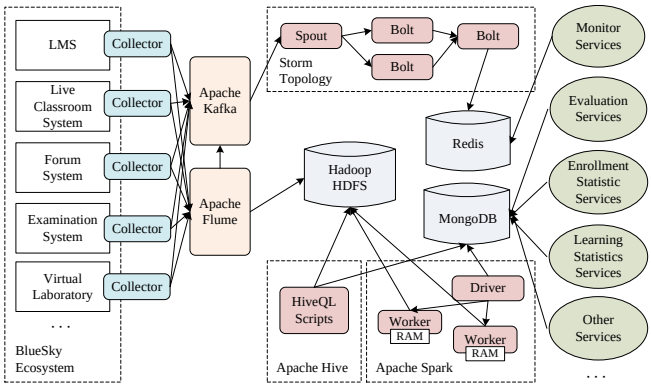

Figure. 2. An implemented logging architecture based on BlueSky ecosystem

A. *Computing students' admission and attendance situations*

According to the scheduled curriculum, all students should login LMS (Learning Management System) to attend class in each semester. However, some students fail to attend class in time for various reasons. In order to ensure these students to be notified to avoid delaying learning, students' admission and attendance situations are in great demands to be analyzed. On the basis of batch processing framework of the Computation Module cooperated with other modules in the logging architecture designed above, the evaluation of students' admission and attendance situations can be implemented as shown in Fig. 2.

Firstly, the log data generated by students in LMS are collected by collector programs, and are transferred to HDFS (Hadoop Distributed File System, an open-source framework for storage and large-scale data processing) by Flume (An open-source software for log transport). Secondly, Hive (An open-source software for batch processing) is used to analyze the log data to obtain statistical analysis results, which are then exported to MongoDB (an open-source document database) and read by the statistic-based/statistical service. Furthermore, the application of the statistical service described above is illustrated as follow.

According to a specifically implemented attendance statistical service which collected 17 GB (15,489,655 rows) original log data corresponding to the behaviors of students in LMS, the statistical analysis results of log data that was compressed into 1.2 GB and analyzed by Hive are shown in Fig. 3, where the horizontal axes shows date (in units of per day), and the vertical axes shows the number of students attending class. In spring of 2014, 10,890 students were enrolled for BlueSky platform. They were asked to attend online classes since Mar. 25 2014. Nevertheless, on Apr. 14 2014, with the assistance of attendance statistical service provided in logging architecture, staff of BlueSky platform observed that only 70.82% of the enrolled students had attended their classes while hadn't the rest who were sequentially notified by Email, SMS or phone calls to. Consequently, an obvious increase of the number of students attending class appeared in Fig. 3. Furthermore, on Apr. 29

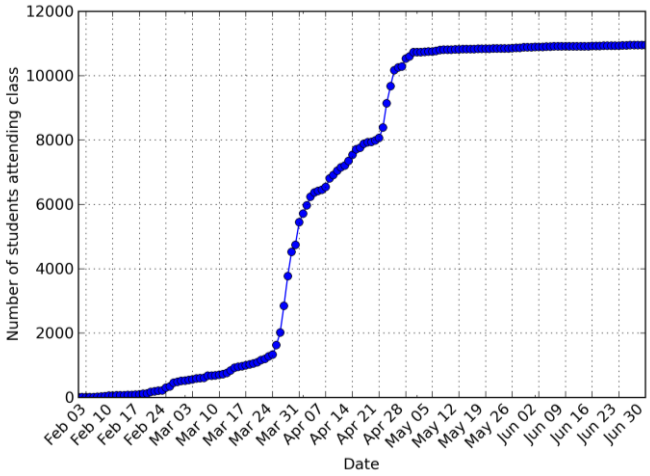

Figure. 3. Statistics for number of students attending class

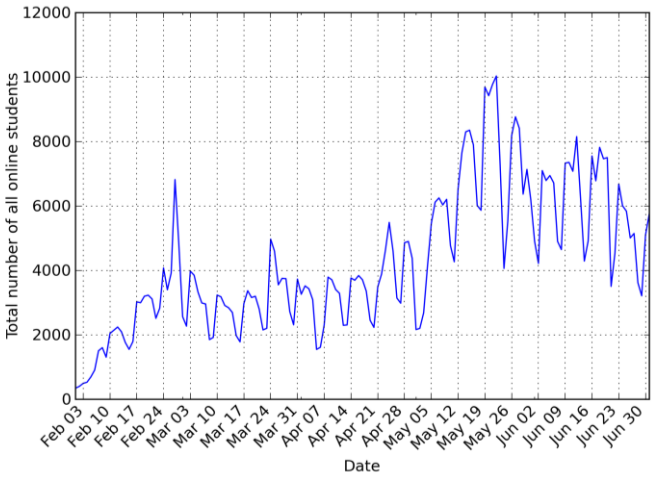

Figure.4. Statistics for total number of all online students

2014, 97.37% of enrolled students had actually attended their classes since being notified.

Furthermore, statistical analysis results of the number of online learners, one aspect of which is the statistical total number of all online students per day shown in Fig. 4, are provided by online attendance statistical service for researchers and data center administrators.

As shown in Fig 4, where the horizontal axis shows date (in units of per day), and the vertical axes shows the total number of all online students, the curve reveals obvious periodicity. Since each column in Fig. 4 represents a week, on the first day of weekdays the curve increases obviously. In addition, on the weekend of each week or holidays, the total number of online students decreased significantly. This phenomenon is resulted by many factors such as learning styles and curriculum plan etc. Therefore, more services including query service, index service and visualization service etc., are expected to analyze detailed statistics of log data, which will be discussed in future work.

B. Live Classroom monitor

Live Classroom is an important system on BlueSky platform. Students can attend practical real-time online lecture in this system and communicate with teachers in text or voice. Furthermore, teachers can arrange quiz or virtual experiments in class. During this teaching process, teachers require the statistics of learning states of online students, discussions, test scores and experiment results to adjust teaching schedule to improve learning performance of students.

The streaming computing framework of Computation Module is applied to satisfy the demands described above, the process of which is shown in Fig 2. Firstly, the log data of students' attending online lecture, discussions in class, test scores and experiment results are collected and transferred to a spout in a topology in storm by Kafka (Storm is an open-source distributed stream computing framework. Topology is a concept defined in Storm and represents a graph of data processing. Spout is a primitive defined in Storm as a data source. Kafka is an open-source high-throughput distributed messaging system). Secondly, the statistics are computed by bolts (Bolt is a primitive defined in Storm as a computing unit) in certain topologies, and the statistical results are stored in Redis (an open-source high-performance in-memory database). Finally, the monitor service reads statistic results from Redis and provides for teachers.

In general, based on Storm, Kakfa and Redis, real-time statistical analysis results of online learning information are provided for teachers to improve teaching performance.

C. Evaluation of student attitudinal performance

On BlueSky platform, the learning outcomes of students consist of academic and attitudinal performances. The academic performance includes scores of examinations and homework which are easily assessed. Whereas it is difficult to evaluate the attitudinal performance. In traditional class education, teachers can evaluate the attitudinal performance of students according to their learning activities and responses towards teaching process in class, which is comparatively difficult in e-Learning environment due to students being out of direct contact with teachers in front of a screen, the large number of online students and the complexity of large-scale learning log data. Therefore, an analysis framework of students' online operations is urgently needed to evaluate student attitudinal performance on the basis of log data.

In the Computation Module, the iterative computing framework which supports implementations of complex data mining algorithms on massive log data satisfies the computation demands. The application process of the iterative computing framework is shown in Fig. 2. Firstly, the log data of video server, LMS, Live Classroom and forum etc., in each layer of BlueSky ecosystem are collected and transferred by Flume to HDFS. Secondly, a certain data mining algorithm is computed by Spark (an open-source cluster analytics framework which supports iterative computing), and the results are saved in MongoDB. Finally, the evaluation service reads results from MongoDB and provides student attitudinal performance for teachers.

Based on the Spark-based iterative computing framework, the demands of teachers for evaluating student attitudinal performance could be satisfied effectively and efficiently. The detailed process will be analyzed in future work.

VI. Conclusion and Future Work

The log data and log analysis are both important bases of e-Learning ecosystem, which indicates that a logging architecture is needed to support the whole life cycle of log data. In this paper, a logging architecture composed of five modules including log collection, transport, storage, computation and service is designed according to the characteristics of log data and log analysis in e-Learning ecosystem, which aims at satisfying the demands of different roles in each layer of e-Learning ecosystem. In addition, an implementation of the proposed logging architecture is developed on the basis of a realistic e-Learning ecosystem. To verify the proposed logging architecture, three typical scenarios are carried out, in which certain log data processing with different computation frameworks are described. Future work will involve more studies on applying the proposed logging architecture to analyze the learning patterns of online learners, and to evaluate learning performance in e-Learning ecosystem.

Acknowledgment

This research was partially supported by National Science Foundation of China under Grant Nos. 91118005, 91218301, 61103160 and 61103239, Key Projects in the National Science and Technology Pillar Program under Grant No. 2012BAH16F02, the Ministry of Education Innovation Research Team No. IRT13035, the MOE-Intel Special Research Foundation of Information Technology under Grant No. MOE-INTEL-2012-04, and the Creative Project in Shanghai Science and Technology Council under Grant No.12dz1506200.

References

[1] J. M. Liss, "Creative destruction and globalization: The rise of massive standardized education platforms," Globalizations vol.10(4), 2013, pp. 557-570.

[2] E. Teall, M. Wang, V. Callaghan, and J.W.P. Ng, "An Exposition of Current Mobile Learning Design Guidelines and Frameworks," International Journal on E-Learning vol.13(1), 2014, pp. 79-99.

[3] L. Breslow, D. E. Pritchard, J. DeBoer, G. S. Stump, A. D. Ho, and D. T. Seaton. "Studying learning in the worldwide classroom: Research into edX's first MOOC," Research & Practice in Assessment, vol.8, 2013, pp: 13-25.

[4] T. KARSENTI, "What the research says," International Journal of Technologies in Higher Education, vol.10(2), 2013, pp. 23-37.

[5] L.Uden, I. T. Wangsa, and E. Damiani. "The future of E-Learning: E-Learning ecosystem," in Digital EcoSystems and Technologies Conference, DEST'07. Inaugural IEEE-IES, IEEE, 2007, pp. 113-117.

[6] B. Dong, Q. H. Zheng, J. Yang, H. Li, and M. Qiao, "An e-Learning ecosystem based on cloud computing infrastructure," in Proc. 9th IEEE International Conference on Advanced Learning Technologies (ICALT), IEEE, 2009. pp. 125-127.

[7] J. H. Lee, and A. Segev, "Knowledge maps for e-Learning," Computers & Education, vol.59(2), 2012, pp. 353-364.

[8] B. Dong, Q. H. Zheng, M. Qiao, J. Shu, and J. Yang. "BlueSky cloud framework: an e-Learning framework embracing cloud computing." In Cloud Computing, Springer Berlin Heidelberg, 2009, pp. 577-582.

[9] A. Fernández, D. Peralta, J. M. Benítez, and F. Herrera, "E-Learning and educational data mining in cloud computing: an overview," International Journal of Learning Technology vol.9(1), 2014, pp. 25-52.

[10] C. Romero, and S. Ventura, "Educational data mining: a review of the state of the art," IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol.40(6), 2010, pp. 601-618.

[11] B. S. Jong, T. Y. Chan, and Y.L. Wu, "Learning Log Explorer in E-Learning Diagnosis," IEEE Transactions on education, vol.16(3), 2007, pp. 216-228.

[12] M.Capay, M. Mesarosova, and Z. Balogh, "Analysis of students' behaviour in e-Learning system," in Proceedings of the 22nd EAEEIE Annual Conference (EAEEIE), IEEE, 2011, pp. 1-6.

[13] R.Mahajan, J. S. Sodhi, and V. Mahajan, "Mining user access patterns efficiently for adaptive e-Learning environment, " International Journal of e-Education, e-Business, e-Management and e-Learning(IJEEEE), vol.2, 2012, pp. 277-279.

[14] A. Kleftodimos, and E. Georgios, "A framework for recording, monitoring and analyzing learner behavior while watching and interacting with online educational videos," in IEEE 13th International Conference on Advanced Learning Technologies (ICALT), IEEE, 2013, pp. 20-22.

[15] A. Klašnja-Milićević, B. Vesin, M. Ivanović, and Z. Budimac, "E-Learning personalization based on hybrid recommendation strategy and learning style identification," Computers & Education vol.56(3), 2011, pp. 885-899.

[16] A. Thusoo, Z. Shao, S. Anthony, D. Borthakur, N. Jain, J. S. Sarma, R. Murthy, and H. Liu, "Data warehousing and analytics infrastructure at facebook," in Proc. ACM SIGMOD International Conference on Management of data, ACM, 2010, pp. 1013–1020.

[17] L. Abraham, J. Allen, O. Barykin, V. Borkar, B. Chopra, C. Gerea, D. Merl et al, "Scuba: diving into data at facebook," in 39th International Conference on Very Large Data Bases, vol. 6, 2013, pp. 1057-1067.

[18] G. Lee, J.Lin, C.Liu, A.Lorek, and D.Ryaboy, "The unified logging infrastructure for data analytics at twitter," in Proc. 38th Proceedings of the VLDB Endowment, vol.5(12), 2012, pp. 1771-1780.

[19] J.Lin, and D.Ryaboy, "Scaling big data mining infrastructure: the twitter experience," ACM SIGKDD Explorations Newsletter, vol. 14, 2013, pp. 6-19.

[20] G. Mishne, J.Dalton, Z. Li, A. Sharma, and J. Lin, "Fast data in the era of big data: Twitter's real-time related query suggestion architecture," in Proc. 2013 International Conference on Management of data, ACM, 2013, pp. 1147-1158.

[21] R.Sumbaly, J.Kreps, and S.Shah, "The big data ecosystem at linkedin," in Proc. 2013 International Conference on Management of data, 2013, pp.1125-1134

[22] Xian Jiaotong University Distance Learning College website, accessible at: www.xjtudlc.com