Online Education Big Data Platform

Guigang Zhang[1*],Yi Yang[1], Xiaoshuang Zhai[1,2], Qi Yao[1], Jian Wang[1*]
1. Institute of Automation, Chinese Academy of Science, Beijing, China
2. Department of automation, Beijing Institute of Technology, Beijing, China
{guigang.zhang, jian.wang, yangyi}@ia.ac.cn; zhaixiaoshuang90@163.com; yaoqi1228@qq.com
*Corresponding author: Guigang Zhang(guigang.zhang@ia.ac.cn),Jian Wang (jian.wang@ia.ac.cn)

***Abstract*—online education is a popular way for distance education in the world. The data of online education platform is explosively growing. The Big Data technologies can process and analyze the massive data and facilitate the user experience and teaching quality. This paper proposes an online education Big Data platform that aims at improving the quality of education by analyzing a large amount of data generated during the online education using big data technologies. This paper presents the concepts and design of the platform. Finally, an application scenario is showed for describing the use of the platform as well.**

***Keywords-Online learning;Big Data;Online education;Analysis; Cloud Computing; Application;* .**

I. INTRODUCTION

Nowadays, online education or e-learning is a popular way to obtain new knowledge for a better career or interests. Online education provides educational resources by Internet so that users can study anywhere and at any time. The data related to the online education e.g., users' learning behavior are worth analyzing for identifying patterns in order to facilitate the teaching quality and user experience of online education. The data are generated rapidly and the data volume is very large. In this case, the traditional method cannot effectively deal with the massive data, let alone analysis. In order to address this issue, we introduce big data technologies to online education system. In this paper, we will discuss the Online Education Big Data Platform with respect to the following points:

- Concepts and architecture of the platform
- Modules of the platform
- Design of the platform

The rest of the paper is organized as follows. Section 2 summarizes the related research work regarding the online education and big data technologies. Section 3 introduces the Online Education Big Data Platform including the architecture and the functional modules. Section 4 shows the design of the public cloud center of the platform. Section 5 presents examples of the proposed Platform. Finally, conclusion and future work are given in Section 6.

II. RELATED WORK

A. Online Education

In the past twenty years, online education has gained much popularity and grown as a main part of the education field. Ch and Popuri [1] made a study on online learning platforms and edX. They found the rise of edX as a global learning platform which has taken online education to a new level. Nowadays, researchers have made studies of online education from different aspects, such as platform, technology, quality. Shanshan [3] designed an online education interaction platform for home-school cooperation. On this platform, students and teachers can exchange information in time, search and publish related information resources, or discuss online. Because there are no manually students behavior monitoring by online education mechanism, researchers have come up with various ideas to measure the student behaviors. Jayasinghe et al. [4] made a comparison among the various proposed ideas aiming at performance evaluation of the students engaged in online education systems. Ahmad et al. [5] focused on the factor impacts to the quality of online education. Du et al [6] presented an online education ontology model in five-tuple in order to solve the unambiguous and uniform expression of online education. Sun et al. [7] proposed an online education approach that can provide teachers and students the online synchronized education and real-time interactions using web operation record and replay techniques. Tu and Liu [8] introduced an autonomous online education system based on intelligent recommendation to meet the high adaptability and individual demand of the learner.

B. Big Data

Big data platform is a system that basically consists of storage component and computing component. The most famous big data platforms are Google cloud platform [9] and Hadoop ecosystem [10] [11]. Google cloud platform contains three core parts: GFS [12], MapReduce [13], and BigTable [14]. GFS (Google File System) is a scalable, high available, distributed file system which works on massive data store. MapReduce is a parallel programming model for batch data processing of large-scale data set. BigTable is a distributed big data storage system based on GFS. Apache Hadoop is a widely used open source big data platform. The core parts include Hadoop MapReduce and HDFS, a scalable and high available distributed system. Moreover, Hadoop ecosystem contains third part components. Apache HBase [15] is a columnar database. Apache Hive [16] is a data warehouse based on Hadoop, which supports data query with SQL-like statements. Apache Spark [17] [18] [19] [20]is an in-memory computing framework that carries out MapReduce computing fully in memory for reducing the Disk I/O cost of the traditional MapReduce methods. Apache Spark framework is growing

rapidly these days because it shows significant advantages than the MapReduce model.

III. ONLINE EDUCATON BIG DATA PLATFORM

In order to manage and analyze the massive educational data, we propose an Online Education Big Data Platform. This section will introduce the architecture and platform functional modules of the platform.

A. *Architechture*

The application scenario of online education big data platform is showed in Figure 1. The platform contains four basic parts: education raw data, data collection/cleaning, educational big data store, and big data analytics for different kinds of users.

1) Educational raw data

The raw data of online education include Student information, Teacher information, Study materials, Teaching materials, Homework, Exam materials, PPT Slides, Study history, Test history, Study notes, Information of other communities, Correction history, Teaching video, as well as BBS comments.

2) Data collection/cleaning

This part works on collecting and cleansing the educational raw data for the future data analysis.

3) Educational Big data store

After collecting and cleansing the educational raw data, the educational data can be stored in three different storage systems. The structured data such as student information and teacher information is stored in MySQL; the semi-structured data such as textual comments in BBS and index data are stored in HBase; the unstructured data like PPT slides and videos are stored with the distributed file system HDFS.

4) Big data analytics for different kinds of users

The educational big data stored can be used for various applications such as computer educational software and mobile Apps. For example, a student can figure out the problems against improving her school record. And, the big data analytics can help a student make her suitable study plan. For a teacher, the big data analytics can estimate the advantages and disadvantages of her teaching and suggest the improvement methods. Meanwhile, big data analytics for education can support the president and the dean to make decisions, for example, the new classes that are needed, and the methods for improving teaching quality.

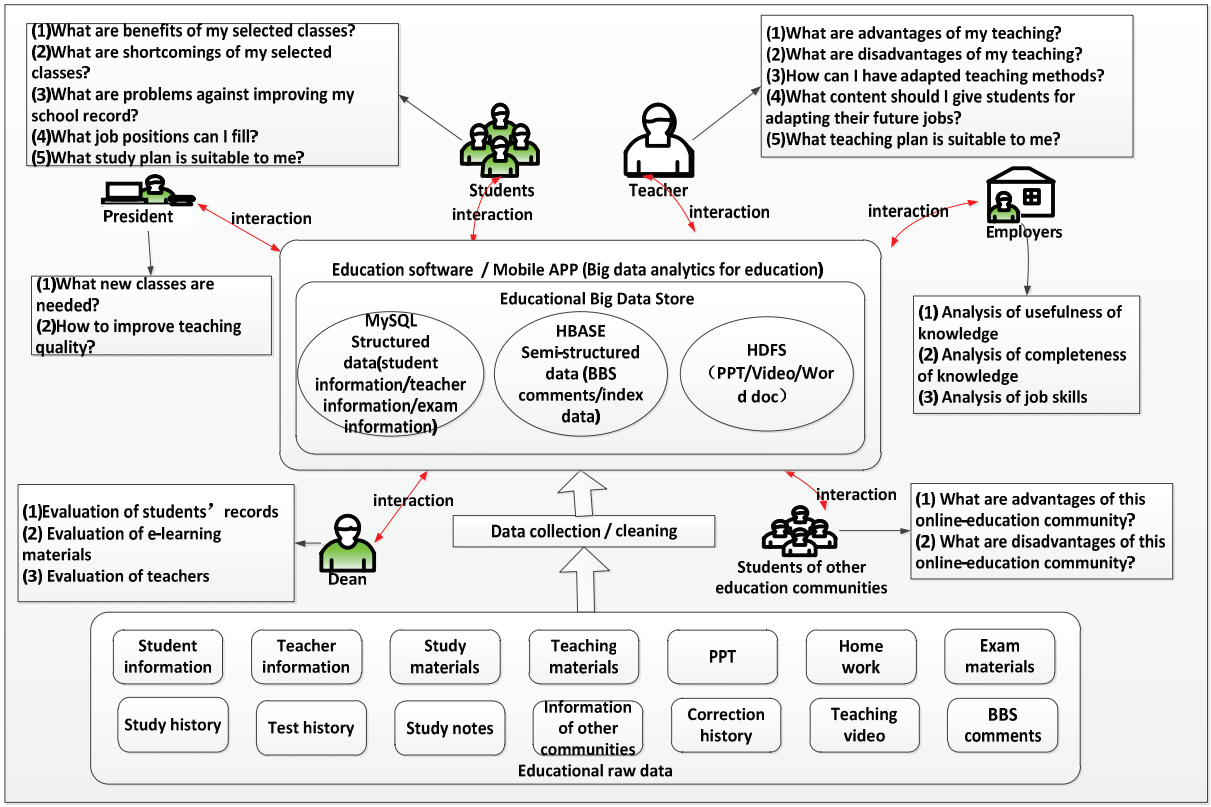

Figure 1 Architecture of the Online Education Big Data Platform

B. *Platform*

The platform is available by a couple of functional modules. The platform is showed in Figure 2. There are three core parts, namely online educational resource collection platform, online educational e-business cloud platform, and online education service platform.

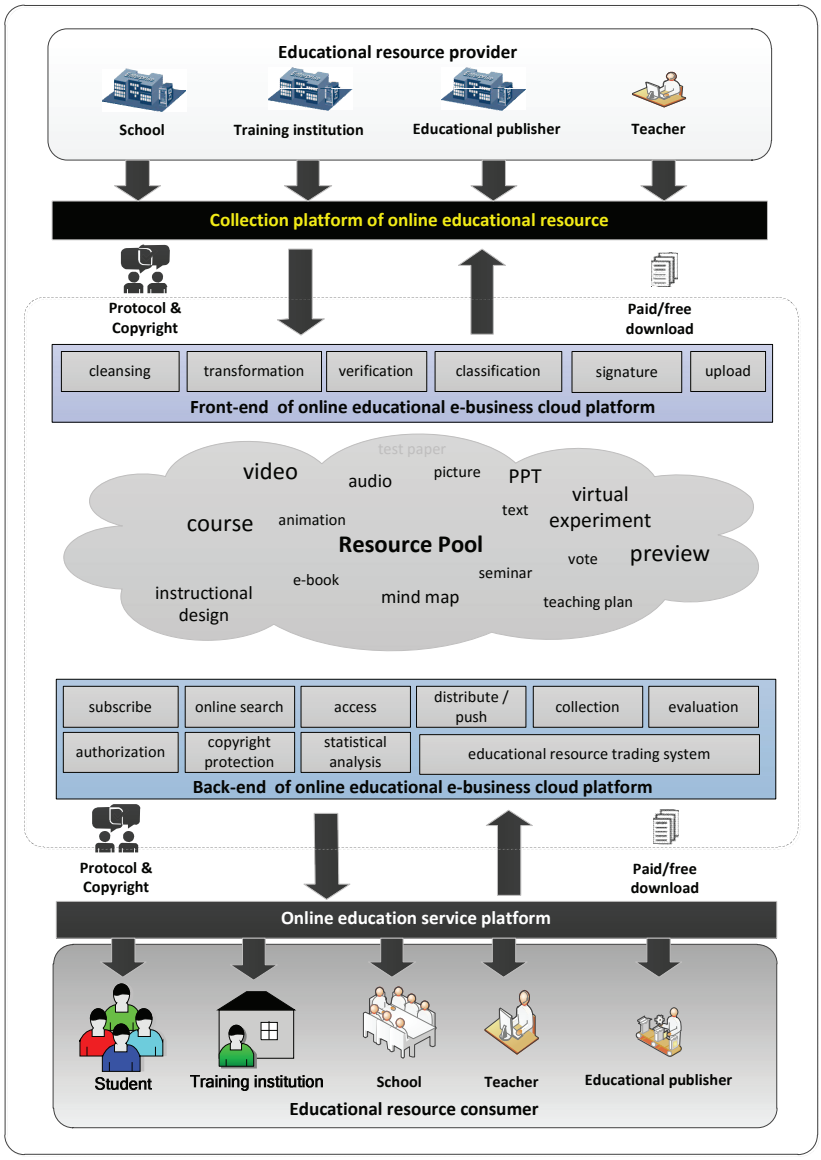

Figure 2 Functional Modules of the Online Education Big Data Platform

1) Collection Platform

The collection platform collects resources from educational resource provider, including schools, teachers, training institutions, and educational publishers. Educational resources mainly include various kinds of materials such as course, video, audio, PPT slide, e-book, virtual experiment. Educational resource consumers basically are schools, teachers, training institutions, educational publishers, as well as students.

2) Online Educational E-Business Cloud Platform

This part contains two basic components: Front-end of cloud platform and Back-end cloud platform.

- The Front-end of cloud platform works on data collection and management.
  - Cleaning: remove duplicate information, correct the error, and provide data consistency.
  - Transformation: converting data format into a suitable format
  - Verification: withdraw data if it does not conform to the standard
  - Classification: classify the resources according to the format
  - Signature: generate digital signatures and encryption for resources
  - Upload: finally deposit the resource into the pool.
- The Back-end cloud platform provides basic modules for supporting the services platform.
  - Subscribe: subscribe to relevant resources of interest.
  - Online search: search for resource.

- Access: interface for data access.
- Distribute/push: according to the demand of the purchaser, the corresponding digital resources are provided.
- Collection: collect users' feedbacks.
- Evaluation: evaluate recommendation accuracy.
- Authorization: assign users permissions.
- Copyright protection: transmit the encrypted file to prevent the leakage of content.
- Statistical analysis: analyze the trend of resources, the proportion of all kinds of resources in a period of time. Provide a report generation tool that can export or print the data report.

3) Online Education Service Platform

This part works on providing various kinds of services for users by using the functional modules in Back-end cloud platform.

IV. Public Cloud Center of the Platform Design

In order to implement the Online Education Big Data Platform, we need a cloud center as a basic platform. We design the Online Education Big Data Platform according to the standard cloud computing architecture. The architecture of the public cloud center is shown in Figure 3.

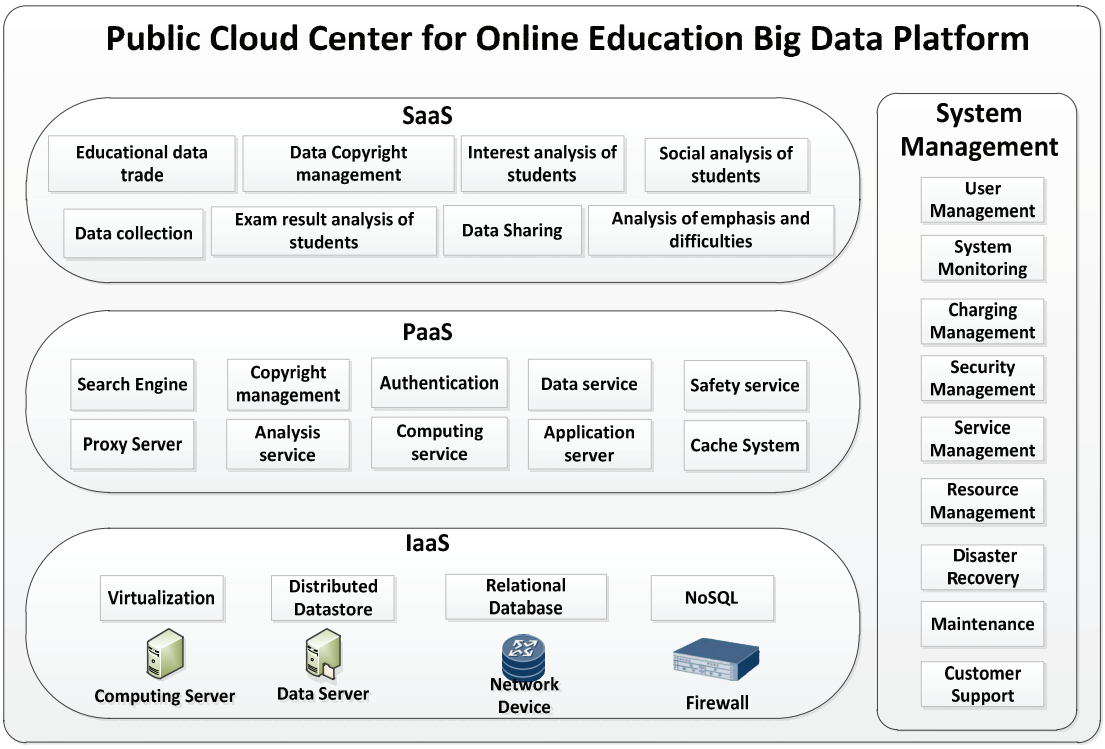

Figure 3 Public Cloud Center of the Online Education Big Data Platform

A. The IaaS Layer

The IaaS layer focuses on the infrastructure of the platform. This layer contains the following parts:

4) Hardware Resource

- Computing server: supports the computation of the platform and the attached applications.
- Data Server: provides the physical resource for data store service.
- Network Device: support the network of the platform.
- Firewall Device: focuses on the security of the platform.

5) Software Resource

- Virtualization: supports to build virtual resources, such as virtual computer cluster
- Distributed data store: works on storing the non-structured master data, e.g., course video.
- Relational database: works on storing the structured data, such as the meta-data of the master data and the user data.
- NoSQL database: focuses on storing the fast-growing data, e.g., logs and users' behavior data.

B. The PaaS Layer

The PaaS layer provides platform services and analysis system services. Most of the modules in the PaaS layer

correspond to the functional modules of the Online Education Big Data Platform introduced in *Section III*.

- Search engine: for online search
- Copyright Management: for copyright protection
- Authentication: for permission control
- Data service: for data access
- Safety service: for data safety
- Proxy server: provides proxy and reverse proxy as firewall software.
- Analysis service: provides analysis, e.g., statistical analysis.
- Computing service: provides different kinds of computing such parallel computing and streaming computing.
- Application server: supports the application deployment, e.g., Apache Tomcat.
- Cache system: improves the effectiveness of the data access.

C. *The SaaS Layer*

The SaaS layer supports to develop applications. The following applications are core applications of SaaS layer.

- Educational data trade: the e-business of educational data.
- Data Copyright management: Copyright protection of educational data.
- Interest analysis of students: the analysis of interests of students for providing more attractive courses.
- Social analysis of students: analysis of interactions between students for recommending suitable partner of class.
- Data collection: collecting logs and users behavior data.
- Exam result analysis of students: statistical analysis of exams for improving exam quality and teaching quality.
- Data Sharing: interface for sharing data.
- Analysis of emphasis and difficulties of course: analysis of course for improving the course building and teaching quality.

D. *System Management*

- User Management: use registration, user information management.
- System Monitoring: system resource utility monitoring, system performance monitoring.
- Charging Management: fee and cost management.
- Security Management: authorization and authentication.
- Service Management: service registration, service start/restart/stop, service availability.
- Resource Management: (virtual) hardware/software resource management.
- Disaster Recovery: system high availability management, redundancy management.
- Maintenance: system performance optimization.
- Customer Support: customer support information management.

The schools and educational institutions may have their own private cloud platforms that manage their educational resources. The Public Cloud Center communicates with the educational cloud platforms in order to collect and update resources of the private cloud platforms. The communication is showed in Figure 4.

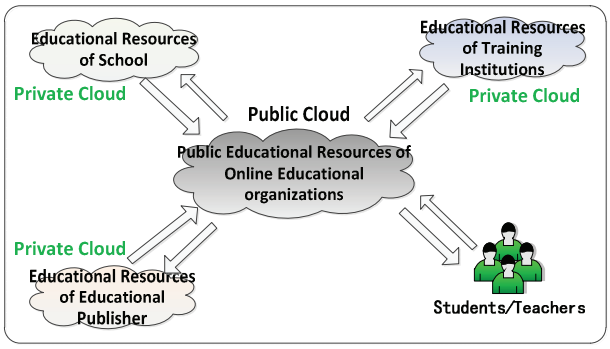

Figure 4 Communication between the Public Cloud Center and private cloud platforms

V. APPLICATION

In this section, we demonstrate examples of our platform for a University for presenting the use of the Online Education Big Data Platform.

For a Student, by collecting and analyzing the student behavior data, e.g., course playback, study notes, the platform can build an interest models for her. This is a very valuable model for different objectives. For the student, the platform can automatically recommend related learning materials and teachers based on the interest model. The platform can also identify the issues of the students in the learning process. On the other hand, teachers may provide tailor courses for the student in order to build good career path the student. In this case the teaching quality will be improved.

Teachers can analyze the learning data of students by using the analysis of the platform. Then, teachers can modify class design according to the analysis results and students' feedbacks in order to improve the attractiveness of the courses and personalized learning suggestions for different students.

The Dean can analyze the hotspots of the online learning by using the platform in order to optimize the course building: introduce new hot courses, remove the outdated courses, new

knowledge training for teachers. Moreover, according to the data analysis of students' comments in BBS or forum, the Dean can extract the opinions of students to courses and teachers as well.

For employers, they can analyze the skills and abilities of the students and make a feasible Recruitment plan for their institutions.

VI. CONCLUSION

In this paper, we have introduces an Online Education Big Data Platform. The platform can integrate the educational big data including educational master data, such as course and homework, user behavior data, and system log data. The platform can share the data using data service and analyze the massive data using machine learning. By using the big data analysis, the platform can facilitate the course building and teaching quality. Meanwhile, the platform provides a bridge between students and job providers in order to help students for better career. The real-time data analysis will be more significant for online analysis of online education so that in the future, we will focus on the streaming computing for the real-time online education analysis.

ACKNOWLEDGMENT

We would like to thank all colleagues and students who helped for our work and thank the following support: (1) Research of Big Data Collection and Analysis for Public Culture Services (Project of National System Design of Public Cultural Service System 2015/2016) (2) Support Program of the National '12th Five-Year-Plan of China' under Grant No. 2015BAK25B04 and Grant No. 2015BAK25B03; (3) National Cultural Resource Sharing Project (key project of '13th Five-Year Plan of China');

REFERENCES

[1] S. K. Ch and S. Popuri, "Impact of online education: A study on online learning platforms and edX," *MOOC Innovation and Technology in Education (MITE), 2013 IEEE International Conference in*, Jaipur, 2013, pp. 366-370.

[2] edX. https://www.edx.org/ accessed 30-May-2016.

[3] S. Shanshan, "Research and Design of Online Education Interaction Platform for Home-School Cooperation," *Intelligent Systems Design and Engineering Applications (ISDEA), 2014 Fifth International Conference on*, Hunan, 2014, pp. 735-739.

[4] U. Jayasinghe, A. Dharmaratne and A. Atukorale, "Students' performance evaluation in online education system Vs traditional education system," *Remote Engineering and Virtual Instrumentation (REV), 2015 12th International Conference on*, Bangkok, 2015, pp. 131-135.

[5] A.Ahmad, I. Naqvi and K. u. Rehman, "Quality of Training Material for Student Learning in Online Education System," *2009 International Conference on Education Technology and Computer*, Singapore, 2009, pp. 316-320.

[6] L. Du, G. Zheng, B. You, L. Bai and X. Zhang, "Research of Online Education Ontology Model," *Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on*, Chongqing, 2012, pp. 780-783.

[7] Y. Sun, D. Chen, W. Jiao and G. Huang, "An Online Education Approach Using Web Operation Record and Replay Techniques," *Computer Software and Applications Conference (COMPSAC), 2014 IEEE 38th Annual*, Vasteras, 2014, pp. 456-465.

[8] Qingsong Tu and Jian Liu, "Research on autonomous online education system based on intelligent recommendation," *IT in Medicine and Education (ITME), 2011 International Symposium on*, Cuangzhou, 2011, pp. 410-413.

[9] Google Cloud Computing. https://cloud.google.com accessed 01-May-2016.

[10] Apache Hadoop. http://hadoop.apache.org/ accessed 01-May-2016.

[11] K.Shvachko, K.H. Rong,S.Radia,et al. The Hadoop Distributed File System: Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium, 2010. Incline Village, NV:IEEE, 2010:1-10.

[12] J. Venner. Pro Hadoop. Apress, 2009.

[13] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, Dec. 2004.

[14] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber. Bigtable: A Distributed Storage System for Structured Data. OSDI'06: Seventh Symposium on Operating System Design and Implementation, Seattle, WA, November, 2006.

[15] Apache HBase. http://hbase.apache.org/ accessed 01-May-2016.

[16] Apache Hive. https://hive.apache.org/ accessed 01-May-2016.

[17] Apache Spark. http://spark.apache.org accessed 01-May-2016.

[18] M. Zaharia, T. Das, H.Y. Li, et al. Discretized Streams: An Efficient and Fault-Tolerant Model for Stream Processing on Large Clusters. Proceedings of the 4th USENIX conference on Hot Topics in Cloud Ccomputing, Pages 10-10, USENIX Association Berkeley, CA, USA, 2012.

[19] Michael Armbrust, Reynold S. Xin, Cheng Lian. Spark SQL: Relational Data Processing in Spark. SIGMOD '15 Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Pages 1383-1394, ACM New York, NY, USA 2015.

[20] Reynold S. Xin, Joseph E. Gonzalez, Michael J. Franklin, et al. GraphX: a resilient distributed graph system on Spark. GRADES '13 First International Workshop on Graph Data Management Experiences and Systems, Article No. 2. ACM New York, NY, USA 2013.