Cloud curriculum resource management platform based on Hadoop

Yu Zhao and Hongxin Liu

Abstract
The rapid development of online education, knowledge sharing, big data, and artificial intelligence technology has brought innovation to education. With the popularization of online education, the variety of teaching resources and the amount of curriculum resources have exploded, and traditional curriculum platforms have been unable to meet the growing demand. The paper builds a cloud computing platform based on Hadoop, which is used for big data analysis and application curriculum resource management. It elaborates on the architecture, functional design, resource storage, and implementation and test of the cloud resource platform. The cloud platform applies to and serves the field of education, which improves resource utilization and sharing, and it can provide better services for teachers, students, and staff.

Keywords
Cloud computing, cloud platform, curriculum resource management

Date received: 23 May 2020; accepted: 17 July 2020

Introduction
At present, education is combined with computer science increasingly, especially with big data, cloud computing, and artificial intelligence technologies. Through the big data analysis to optimize the educational mechanism and allocate educational resources, it can make more scientific decisions, whether educational management departments or teachers, or students, can get personalized applications for different users, which is a potential educational revolution. Big data networking exchange demand promotes the upgrading and expansion of IT infrastructure. Social media, the mobile Internet, and the Internet of things (IOT) have allowed data to grow exponentially over the past decade, with data generation rates, diversity, and complexity increasing. With the development of massive open online courses (MOOCs) and online education, digital teaching resources such as teaching courseware, images, video, and audio explode, and the existing online teaching environment is insufficient to meet the increasing demand for data access. The curriculum resource platform can store a large amount of teaching resources. In the process of using the traditional curriculum resource platform, the teaching resources accumulate more and more, and the pressure on the resource online service system is increasing. Meanwhile, it affects the access of the platform. The challenge is to build a platform to meet the resource management needs, which can solve the contradiction between insufficient resource storage, insufficient access speed, and imperfect storage structure, and utilize the ability of the cluster to reduce the response time of the accessing system.

Amazon and Google launched their cloud computing services in 2006, and Microsoft launched Azure cloud computing services in 2008. Amazon’s AWS (Amazon Web Services) accounted for most of the public cloud computing area. As an example of AWS EC2 cloud hosting services, new product lines are launched every year. Cloud computing is a computing model that provides resources through the Internet in a service manner. This method is dynamic and scalable, and can integrate various types of resources in the network on the cloud platform to work together, which is a service way to provide access to business and storage functions. Millions of users and IOT devices have connected to the cloud platform, and Apps are numerous. There are 12 elements to build and run cloud applications: codebase with multiple deployments, clear dependencies, storage configuration in the environment, and...
back-end services as additional resources, strictly separating build and run, process stateless and unshareable, service provided through port binding, extension process model, disposability, development environment is equivalent to online environment, log is event flow, and background management task is one-time process. The online education platform built by cloud computing can achieve the goals of education fairness, educational achievement sharing, efficient and convenient learning, and lower learning costs; enrich teaching forms; and enhance teaching effects. The realization of these goals depends on the emergence of the cloud concept and technology. When the cloud computing platform is used in the field of education, it can provide effective services for teaching, experimentation, after-school learning, and after-school tutoring. The cloud curriculum platform can improve teachers’ teaching efficiency, promote students’ interactive learning, improve students’ thinking ability, promote the development of student group wisdom, and improve the overall teaching quality. The adoption of the cloud platform is conducive to the reform of the teaching model, breaking the tedious and boring traditional teaching and enhancing delight in study. The course of Big Data Analysis and Application includes systematic understanding and practical application of the whole process of big data collection, storage, calculation, processing, analysis, mining, and visualization. The course focuses on cultivating students’ ability to apply theory to practice. In the class, teachers explain the relevant concepts and principles in big data analysis to students and develop students’ ability to analyze big data through the curriculum design experiment course. GFS (Google File System) is a distributed file system designed by Google to improve the efficiency of distributed storage and access to data in large-scale PC clusters. Google released MapReduce again in 2004, describing a programming model for data processing and generation of large-scale clusters. These two papers inspired Yahoo to invite Doug Cutting in 2004 to develop Hadoop, which is the well-known open-source distributed data storage and processing software architecture. Therefore, the paper uses Hadoop to build a resource management platform that can not only enable students to master the big data development platform, but also concentrate the big data curriculum resources into the cloud to realize the sharing of curriculum resources. This paper constructs a cloud computing model first and then builds a cloud platform based on Hadoop distributed frame. It describes the platform design, data storage, and performance test in detail.

In summary, most of the studies relied on traditional method to improve online teaching. The work has the following contributions. First, the paper establishes a cloud platform which can integrate hardware resources, improve the utilization of existing hardware resources, reduce hardware costs, and enhance the system access efficiency. Second, the application of cloud computing on the curriculum resource platform improves the construction of teaching resource platform and digital management, and establishes an open, flexible, and interactive teaching resource service platform, which is beneficial to build a communication platform between teachers and students. Third, cloud computing realizes the sharing of teaching resources, teachers can get more excellent teaching materials and experience, and students can share learning results and experience, which is beneficial to the dissemination of knowledge. Finally, the research improves the way of resource storage and access, and is convenient to maintain and reproduce.

Related work

Google and IBM have established multiple cloud computing data service centers to provide cloud computing services to universities around the world. The University of Westminster adopted the Google platform which provided services of email, word processing, spreadsheet, presentation, and group-based assignments. The cloud computing platform was adopted by schools of the Kentucky’s Pike County district, which was developed by IBM. Microsoft helped Ethiopia construct education system through Microsoft’s Azure cloud platform. Teachers can use and manage cloud resources to work on the cloud and create tools. They all used cloud services provided by big companies. The research trend of cloud computing focused on the technology, application, cost, benefit, and security of enterprises and is less in the field of education.

There are some works listed, which show the application successful business model of cloud computing to education area. Customer relationship management (CRM) model was used to manage the relationship between student and staff. The model was based on the cloud computing infrastructure system. Chang and Wills elaborated on how to apply the supply chain business model to build an education cloud, and it achieved preferable result in a university. Pardeshi proposed deployment models and service model of cloud computing for higher education, and provided recommendation for migration from the traditional system to cloud computing systems. Kalagiakos and Karampelias presented a computing ecosystem applicable to the cloud education and suggested that API (Application Programming Interface) standards and learning resources should be defined clearly. Sabi et al. explored the application of a cloud computing model considering background, economic, and technological implications to universities in developing countries. Koch et al. established cost-optimization mechanism, designed resource allocation plan, and improved the utilization of computing platform in educational environment application.

Many works are in progress to build a virtual laboratory or virtual learning environment based on cloud computing.
computing frame that enabled teachers to deploy customized experiment environments. Others developed learning management system to support learning, based on the cloud computing platform. The virtual computing environment was established for computer networks course. A North Carolina rural high school used an educational cloud to access algebra software. Caminero et al. used cloud computing to build a virtual remote laboratory, which was not simulation and does not need to access hardware.

There are also some tools to analyze students’ behavior, to perform personalized learning and resource recommendations. Knewton used big data to provide adaptive learning and analytics for students, teachers, and publishers. Knewton analyzed vast amounts of anonymized data to work out what students know and how they can learn better, and then recommended the lessons for students. Ma et al. compared the learning effect with the intelligent tutoring system (ITS) and without the ITS, and proved that ITS was a relatively effective learning tool. Fei and Yeung analyzed the students’ learning behavior on the MOOC platform and predicted the students’ dropout according to their activity. Kavitha built concept map and concept associations to understand resources better and recommended more appropriate resources to users. Arpaci implemented cloud computing in the educational environment to strengthen effective knowledge management, and built and verified the theory model.

Most of the research and application of cloud computing in the field of education are architecture and suggestions for cloud platform, and some of them have not been integrated into a specific platform. There are few cloud computing applications for curriculum resource platform. Based on this, this paper establishes a cloud computing platform which integrates resource management and learning analysis for education actors.

**Method**

The cloud resource platform ecosystem consists of four parts: participation entities, learning resources, teaching activities, and education management. These participants, resources, activities, and services together form a complete environment to enhance the educational and teaching performance, as shown in Figure 1.

The cloud platform is characterized by the migration of students’ learning to cloud, which truly realizes the ubiquity of learning. Compared with the traditional online education platform, it also shows many unique advantages, listed as follows:

1. The application is comprehensive. The cloud education integrates all kinds of teaching software, including teaching, learning, management, communication, and entertainment software.

2. Rich educational resources can be fully utilized and shared equally. A wealth of educational resources is presented to students through the most appropriate organization.

3. Convenient, safe, and cheap. Because the software is stored in cloud, there is no need to download, install, maintain, and upgrade. Only one smart terminal device can be used to learn anytime and anywhere, as long as it can access the Internet. The cloud platform stores data in the cloud and users do not have to worry about data security issues, as long as the cloud has service to manage information.

4. Learn anytime, anywhere. The seamless connection between the cloud platform and mobile terminal devices allows learners to access the platform with simple mobile devices, to obtain learning content and learning methods of interest from the cloud, and to carry out personalized and liberalized mobile learning.
**Cloud curriculum resource platform model**

The big data curriculum resource platform combines cloud computing storage and technology, big data analysis, to make teachers and students understand the current status through data, and it can predict the future status. It can collect, filter, analyze, and mine various data, and the data are suitable for educational decision makers. The curriculum resource model is shown in Figure 2, including application layer, data processing layer, and cloud resource management layer.

The cloud resource management layer mainly includes cloud data storage, cloud resource management, and cloud safety management. The data storage is to save and update the data. The cloud resource management manages the virtual resources of data storage, which monitors, allocates, recovers, updates, and maintains the resources, and carries on the load balance management to each resource node of data storage. The cloud safety management manages user access to each cloud resource, including login verification, authority management, and firewall management.

The data processing layer includes data collection, data analysis, and data mining. Data collection is the data cleaning, data filtering, and data storage of structured, semi-structured, and unstructured data collected in the cloud resource management layer.

The application layer is applied to the processed data, such as information retrieval, data visualization, and teaching and learning evaluation and prediction. Through the evaluation and prediction of teaching and learning, the college departments can improve the efficiency in time. The users of the cloud platform do not directly interact with Hadoop and do not need to program to realize the function of cloud resource management. The users of the cloud platform can query, retrieve, and upload resources through the computer or the tablet using the browser.

**Curriculum resource platform based on Hadoop**

Hadoop is an open-source distributed cloud computing platform of the Apache Software Foundation. Users can use Hadoop to integrate computing resources, build a private distributed cloud computing platform, and use the computing and storage capabilities of the cluster to process massive amounts of data.\(^3^6\) Apache Hadoop solves problems of large amounts of data and computations using a network of many computers, which is famous for HDFS (Hadoop Distributed File System) as storage section and MapReduce programming model.\(^3^7\) The basic Hadoop framework consists of the following modules: Hadoop common, HDFS, Hadoop YARN,\(^3^8\) and Hadoop MapReduce.\(^3^6\) The Hadoop ecosystem is shown in Figure 3, including the additional software of Apache Pig, Hive, Storm, Spark, HBase, ZooKeeper, Flume, Sqoop, and so on.\(^3^9\)

The distributed file system manages files stored across multiple computers in the network. HDFS is a
file system that allows files to be shared across multiple computers through the network, in which multiple users share files and storage on multiple computers. The action of actually accessing the file through the network, as seen by the program and the user, is like accessing a local disk. HDFS has a good fault tolerance mechanism to ensure reliability. Even if some nodes in the system are offline, the system as a whole can continue to operate without data loss. It applies to the case where we run multiple queries and one-time writes.

MapReduce is a parallel programming model proposed by Google in 2004 that can process large amounts of data concurrently. Its parallel computing process is complex and runs on large-scale clusters into Map and Reduce functions. MapReduce solves the analysis and processing of massive data, which is universal, easy to develop, and robust. A MapReduce job typically breaks up the data set into a number of independent data blocks, which are processed by the Map task in a fully parallel manner, and then sends the results into the Reduce task. Generally, the computing nodes and storage nodes of the MapReduce framework run on the same group of nodes, that is, the nodes running MapReduce and HDFS are usually together, which makes the network bandwidth utilization of the cluster more efficient.

HBase is a column-oriented distributed database, which uses HDFS, MapReduce, and ZooKeeper as a collaborative service. It is mainly used for real-time random reading and writing of large-scale data sets.

**Experiment and results**

Based on the distributed framework of Hadoop and data storage of HBase, a cloud curriculum resource platform with reasonable structure is designed. Using powerful storage ability and flexible expansibility of Hadoop, the curriculum resource data can be stored in the resource database, which can not only store teaching resources but also support diversified learning method and optimize the utilization of resources.

From the perspective of education, students, teachers, and education managers are all participants in teaching activities. Using cloud platform and mobile Internet technology, students, teachers, and education managers can achieve direct interaction without time and space constraints, and give full play to the characteristics of each type of entity so that the results of education and teaching benefit each entity. These form a virtuous circle atmosphere.

According to the participant entities in the platform, the use and authority of each entity to resources are standardized. Student users view and modify student information, and use the platform to browse resources, obtain resources, search resources, and provide feedback on resources. Teachers can view the evaluation of teaching resources and modify the resources according to the evaluation. Teachers can view and modify teacher information, manage resources, manage student users, carry out teaching, and publish learning tasks. Administrator can manage user authority, manage user information, audit resources, and publish information. The main attribute of the resource is shown in Figure 4.

**The step of using Hadoop to build a cloud computing platform**

Considering that it is still in the experimental stage and purchasing a group of servers is costly, the paper uses pseudo-distributed installation, which not only saves hardware resources but also enables students to build a Hadoop distributed system in practice.

The virtual machine is installed in the Windows 10 operation system. The CentOS 6.6 Linux system is adopted, loading Hadoop 2.4.1. The cluster consists of...
four virtual machines. Each machine has one CPU and 2G memory. The Hadoop cluster is composed of one master node and three slave nodes.

*Install SSH, configure SSH login without password.* The SSH protocol is a security protocol based on the application layer and transport layer. Hadoop control script relies on SSH to perform operations on the entire cluster, so users need to be allowed to log on to machines in the cluster without password.

It is necessary to create a public/private key pair stored in NFS and let the entire cluster share this key pair.

*Hadoop pseudo-distributed configuration.* Hadoop can run on a single node in a kind of pseudo-distributed manner, which runs as a separate Java process. According to the Hadoop documentation, the files of hadoop-env.sh, yarn-env.sh, hdfs-site.xml, core-site.xml, mapred-site.xml, and yarn-site.xml are configured, which is used to configure the system’s Java installation directory, configure distributed file system HDFS, set HDFS file storage directory, access control permission parameters, configure Hadoop cluster temporary directory, configure log file, configure I/O attributes, and set the data processing and data interaction mode.

*HBase configuration.* The bottom of HBase relies on HDFS, including ZooKeeper, HMaster, and HRegionServer components. HBase needs to configure hbase-env.sh file, regionservers file, and hbase-site.xml file. When hbase-site.xml is configured, specify the access address of HDFS, the working mode of HBase, and HBase metadata storage path. After completing the above configuration of a node, copy other nodes to complete the configuration of the entire cluster. HBase should be started after HDFS startup, and then there is an HBase work directory in HDFS. Users can start HBase Shell, create table.

*Network configuration.* The node in the cluster is copied directly in the virtual machine, so it is necessary to ensure that the MAC (Media Access Control) addresses of each node are different. First, complete the preparation on the master node and close the Hadoop, and then perform subsequent cluster configuration. Modify the host name of each node, transfer the master public key to each slave node, and add authorization to each slave node. Thus, on the master node, it can use SSH to access each slave node without password.

*Resource storage*

The curriculum resources required by the cloud platform are stored in a large-scale cluster of Hadoop, instead of being stored locally, and the storage of the platform is implemented by HDFS. HBase is a distributed non-relational database and is a new type of column-oriented database. The HDFS is used to complete the creation and storage of curriculum resource files, and the distributed storage capability of HDFS is utilized to realize the fast access of resources. Resource storage and read/write interfaces are programmed by Java API from HDFS.

Structured data, such as student information, teacher information, and other data, are stored in the MySQL database and stored in the server's stand-alone node to speed up the query. Unstructured data, such as audio, sound, video, rich text, and images, are stored in the cluster's HBase database. A simple, small amount of structured data stored in MySQL, such as user information and student information, is regularly imported into HDFS for later analysis on the Hadoop platform.

The client obtains the node where the resource is located through the balanced load capability of the

---

**Figure 4. Resource attribute.**
apache server cluster, which can effectively balance the load, control the concurrent of each resource node, and ensure the access speed of the resource.

Platform test

In order to get more understanding of the cloud platform, curriculum resource management platform must be tested. There are many indicators for testing the performance of the Hadoop platform. Usually, the efficiency of the CPU, the throughput of reading and writing files, and the test time are used to evaluate the performance of the Hadoop platform. TestDFSIO test can indirectly test the writing and reading performance of the HDFS of the resource platform. The TestDFSIO program performs concurrent read and write operations through a MapReduce job. Each Map task reads and writes files. The experiment tested six files, which are 100 MB, 1 GB, 2 GB, 5 GB, 10 GB, and 20 GB, respectively. The throughput test results of reading and writing a single file are shown in Figure 5. The execution time of test is shown in Figure 6. It has been observed from the test results that when the amount of data increases, the reading and writing time of the test files shows a steady growth trend in the daily learning situation where the cluster is small and the files are not large. When the number of files increases, the throughput increases, and system status is stable. MRBench test executes a small job repeatedly, which can test the performance of distributed clusters in processing multiple document files. As the number of executions increases, the average time for job execution decreases, shown in Figure 7. From the test results, the performance of the cloud platform increases as the amount of

![Figure 5. Throughput of reading and writing single file.](image1)

![Figure 6. Execution time of reading and writing single file.](image2)
data increases. When the amount of data is small, the efficiency is not high. This article is tested on virtual nodes in virtual machine. If it is widely deployed in the laboratory in the future, the cluster size and node performance will be higher.

**Conclusion**

Education is more intelligent, and the methods of learning and teaching online course are more convenient. The Big Data Curriculum Resource Platform is an intelligent platform based on cloud computing technology, virtualization technology, and distributed storage, which can provide services for different users. The platform can realize intelligent teaching, intelligent learning, and intelligent management and can effectively solve many problems, such as inequality in educational resources and waste of educational resources, realize the on-demand use of resources, and achieve effective sharing of resources.

Based on cloud computing technology, this paper establishes a real-time, interactive, and open big data curriculum resource platform. Data mining technology and learning analysis technology are used to process the data generated by the platform, and the massive data generated by students are interpreted and analyzed to assess student academic progress, predict future performance, and identify potential problems. The platform is applied to the big data analysis and application course of computer science major, and has achieved better teaching and learning results. In the future, it is hoped that the platform will be popularized to more courses and form a better resource sharing platform.

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) received no financial support for the research, authorship, and/or publication of this article.

---

**ORCID iD**

Yu Zhao ORCID: https://orcid.org/0000-0002-0476-3318

**References**

1. https://www.internetlivestats.com/one-second/
2. https://www.ibmbigdatahub.com/infographic/four-vs-big-data
3. El Mhouti A, Erradi M and Nasseh A. Using cloud computing services in e-learning process: benefits and challenges. *Educ Inf Technol* 2018; 23: 893–909.
4. Cusumano M. Cloud computing and SaaS as new computing platforms. *Commun. ACM* 2010; 53: 27–29.
5. MIT Technology Review. https://www.technologyreview.com/2011/10/31/257406/who-coined-cloud-computing/
6. https://aws.amazon.com/blogs/aws/ec2-instance-history/
7. Hashem IAT, Yaqoob I, Anuar NB, et al. The rise of “big data” on cloud computing: review and open research issues. *Inf Syst* 2015; 47: 98–115.
8. Furtht B and Escalante A. *Handbook of cloud computing*. New York: Springer, 2010.
9. IDC’s 3rd platform, https://www.idc.com/promo/thirdplatform
10. https://12factor.net/
11. Baldassarre MT, Cuivano D, Dimauro G, et al. Cloud computing for education: a systematic mapping study. *IEEE Trans Educ* 2018; 61: 1–11.
12. Cappos J, Beschastnikh I, Krishnamurthy A, et al. Seattle: a platform for educational cloud computing. In: *Proceedings of the 40th ACM technical symposium on computer science education*, Chattanooga, TN, 4–7 March 2009, pp. 111–115. New York: ACM.
13. González Martínez-JA, Bote-Lorenzo ML, Gómez-Sánchez E, et al. Cloud computing and education: a state-of-the-art survey. *Comput Educ* 2015; 80: 132–151.
14. Ghemawat S, Gobioff H and Leung ST. The Google file system. *Acm Sigops Operat Syst Rev* 2003; 37: 29.
15. Dean J and Sanjay G. MapReduce: simplified data processing on large clusters. *Commun Acm* 2008; 51: 107–113.
16. White T. *Hadoop: the definitive guide*. Riverstein Highway North, CA: O’Reilly Media Inc, 2012; 215: 1–4.
17. Al-Zoube M, Abou El-Seoud S and Wyne MF. Cloud computing based e-learning system. *Int J Dist Educ Tech* 2010; 8: 58–71.
18. Sultan N. Cloud computing for education: a new dawn? *Int J Inf Manag* 2010; 30: 109–116.
19. https://www.seattletimes.com/business/microsoft/microsoft-cloud-computing-gets-down-to-earth/
20. Chine K. Learning math and statistics on the cloud, towards an EC2-based Google docs-like portal for teaching / learning collaboratively with R and Scilab. In: *Proceedings of the 10th IEEE international conference on advanced learning technologies*, Sousse, Tunisia, 5–7 July 2010.
21. Vulić M, Barač D and Bogdanović Z. CRM as a cloud service in e-education. In: *Proceedings of the 19th telecommunications forum TELFOR* 2011. https://www.researchgate.net/publication/261212636_CRM_as_a_cloud_service_in_e-education
22. Chang V and Wills G. A university of Greenwich case study of cloud computing: education as a service. In: *E-logistics and E-supply chain management: applications for evolving business*. Hershey, PA: IGI Global, 2013, pp. 232–253.
23. Pardeshi VH. Cloud computing for higher education institutes: architecture, strategy and recommendations for effective adaptation. *Procedia Econ Finance* 2014; 11: 589–599.

24. Kalagiakos P and Karampelas P. Cloud computing learning. In: *Proceedings of the 2011 5th international conference on application of information and communication technologies (AICT)*, Baku, Azerbaijan, 12–14 October 2011, pp. 1–4. New York: *IEEE*.

25. Sabi HM, Uzoka F-ME, Langmia K, et al. Conceptualizing a model for adoption of cloud computing in education. *Int J Inf Manag* 2016; 36: 183–191.

26. Koch F, Assunção MD, Cardonha C, et al. Optimising resource costs of cloud computing for education. *Future Gener Comput Syst* 2016; 55: 473–479.

27. Figueiredo RJ, Wolinsky DI and Chuchaisri P. Educational virtual clusters for on-demand MPI/Hadoop/condor in futuregrid. In: *Proceedings of the 2011 TeraGrid conference on extreme digital discovery—TG ’11*, Salt Lake City, Utah, 18–21 July 2011.

28. School of Business. Creating collaborative and convenient learning environment using cloud-based moodle LMS. *Int J Web-based Learn Teach Technol* 2016; 11: 35–50.

29. Dinita RI, Wilson G, Winckles A, et al. A cloud-based virtual computing laboratory for teaching computer networks. In: *Proceedings of the 2012 13th international conference on optimization of electrical and electronic equipment (OPTIM)*, Brasov, 24–26 May 2012.

30. Stein S, Ware J, Laboy J, et al. Improving K-12 pedagogy via a cloud designed for education. *Int J Inf Manag* 2013; 33: 235–241.

31. Caminero AC, Ros S, Hernández R, et al. Virtual remote laboratories management system (tutores): using cloud computing to acquire university practical skills. *IEEE Trans Learn Technol* 2015; 9: 133–145.

32. https://www.knewton.com/

33. Ma W, Adesope OO, Neshit JC, et al. Intelligent tutoring systems and learning outcomes: a meta-analysis. *J Educ Psychol* 2014; 106: 901.

34. Fei M and Yeung D-Y. Temporal models for predicting student dropout in massive open online courses. In: *Proceedings of the 2015 IEEE international conference on data mining workshop (JCDMW)*, Atlantic, NJ, 14–17 November 2015, pp. 256–263. New York: *IEEE*.

35. Arpaci I. Antecedents and consequences of cloud computing adoption in education to achieve knowledge management. *Comput Hum Behav* 2017; 70: 382–390.

36. Apache Hadoop, http://hadoop.apache.org/

37. Shvachko K, Kuang H, Radia S, et al. The Hadoop distributed file system. In: *Proceedings of the 2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*, Incline Village, NV, 3–7 May 2010, pp. 1–10. New York: *IEEE*.

38. Vavilapalli VK, Murthy AC, Douglas C, et al. Apache Hadoop yarn: yet another resource negotiator. In: *Proceedings of the 4th annual symposium on cloud computing*, Santa Clara, CA, 1–3 October 2013, pp. 1–16. New York: *ACM*.

39. https://finance.yahoo.com/news/continuuity-raises-10-million-series-120500471.html

40. Landset S, Khoshgoftaar TM, Richter AN, et al. A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *J Big Data* 2015; 2: 24.