Journal of Education and Learning; Vol. 11, No. 3; 2022
ISSN 1927-5250 E-ISSN 1927-5269
Published by Canadian Center of Science and Education

System Architecture of Big Data in Massive Open Online Courses (BD-MOOCs System Architecture)

Withamon Khajonmote[1], Kittipong Chinsook[2], Sununta Klintawon[1,3], Chaiyan Sakulthai[3], Wicha Leamsakul[3], Natchanok Jansawang[3] & Thada Jantakoon[3]

[1] Department of Computer Technology, Faculty of Science and Technology, Rajabhat Maha Sarakham University, Maha Sarakham, Thailand

[2] Department of Computer Science, Faculty of Science and Technology, Rajabhat Maha Sarakham University, Maha Sarakham, Thailand

[3] Department of Information and Communication Technology for Education, Faculty of Science and Technology, Rajabhat Maha Sarakham University, Maha Sarakham, Thailand

Correspondence: Thada Jantakoon, Department of Information and Communication Technology for Education, Faculty of Science and Technology, Rajabhat Maha Sarakham University, Maha Sarakham, Thailand.

Received: March 4, 2022 Accepted: April 10, 2022 Online Published: April 19, 2022

doi:10.5539/jel.v11n3p105 URL: https://doi.org/10.5539/jel.v11n3p105

Abstract

The system architecture of big data in massive open online courses (BD-MOOCs System Architecture) is composed of six components. The first component was comprised of big data tools and technologies such as Hadoop, YARN, HDFS, Spark, Hive, Sqoop, and Flume. The second component was educational data science, which is composed of the following four parts: EDM, ERS, AA, and S/II. The third component was a description of three basic elements of a big data system: data capture, management, and analysis. The fourth component was that MOOCs were classified as cMOOCs, xMOOCs, quasi-MOOCs, hMOOCs, and other related. The fifth component included the steps of MOOC development: design, delivery, and assessment. Finally, MOOCs present educational data science challenges such as analyzing student interactions, estimating dropout risk, grading, and making recommendations. Overall, the BD-MOOCs system architecture design was suitable at the highest level.

Keywords: big data, data science, bd-moocs, predicting students, big data in mooc, learning recommendations

1. Introduction

MOOCs have been hailed as the perfect educational research platform. Their massive sample sizes and capacity to monitor precise student engagement throughout the course provide unequaled opportunities to conduct learning experiments (Romero & Ventura, 2016). This gold mine has resulted in an increase in research investigations and the application of EDS methods to the data collected on these platforms. A quality education is essential for long-term social and economic development as well as cross-cultural understanding, according to UNESCO. Open Educational Resources (OER) provide a strategic opportunity to increase educational quality while also facilitating policy discourse, information exchange, and capacity development (UNESCO, 2002). MOOCs, or Massive Online Open Courses, are a modern e-learning trend. According to Siemens (Toth, 2019), MOOCs are a continuation of the trend in innovation, experimentation, and the use of technology that distance and online learning pioneered to give learning opportunities to vast numbers of learners. According to the Commonwealth of Learning (Kinskey, King, & Lewis, 2018), MOOCs are a means of facilitating the efficient creation, distribution, and use of knowledge and information for learning by utilizing freely available online resources such as OER. They can be used to support social networking and other forms of “connectivity” among participants. There are a lot of self-organizing learners who take part in the MOOC and connect with each other through the platform and other social interaction tools that are available.

Big data refers to the massive volume of data that typical data management methods cannot handle (Ashabi, Sahibuddin, & Haghighi, 2020). Big data includes the use of technology to process, analyze, and visualize potentially massive datasets on time. Big data has three features: volume, speed, and diversity. Volume refers to the wide range of data produced continually through websites, online education, learning and evaluation of education, and social networking (Kler, 2014). New technologies will address the difficulty of storing this data. The enormous data speed dimension defines how much data is produced. In other words, the speed at which information comes and the time it needs to be acted upon. Existing research analyzes numerous definitions of big data. Big-scale datasets collect large, complex, diversified, and heterogeneous datasets generated by numerous sources such as clickstreams, sensors, sharing videos, economic transactions, and social networks. A comparable definition of big data is large-scale datasets that are beyond the capability of commonly used traditional data management and analytical methodologies to acquire, store, access, manage, exchange, process, analyze, and display in an acceptable amount of time (Chen, Mao, & Liu, 2014; Islam & Reza, 2019). The term “big data” has raced to a new revolution in the data management process of large-scale data in numerous areas of research and technology. As a result, novel tools and approaches in data analytics must be developed to satisfy the difficulties and goals of studying large-scale information. As a result, we demand a framework that maintains high processing speed and intelligent and scalable storage systems for large-scale information. The creation of vast amounts of data presents three key issues. Data volume, data velocity, and a broader range of data types (Kockum & Dacre, 2021). These properties are referred to as the 3V-model (volume, velocity, variety) for big data. Furthermore, the concept extends to a 5V model for describing data qualities such as volume, velocity, variety, veracity, and value (Rao, Mitra, Bhatt, & Goswami, 2018; Saadia, 2021). There is a significant body of literature on the use of big data in education, and this literature is expanding. When processing vast amounts of learning data, it is possible to get insight into the relationship between learning practices and learning efficacy, which can be used to help educators predict learning outcomes (Hwang, Chu, & Yin, 2017). In learning analytics, the conceptual framework that underpins course features, student performance assessment, and prediction of learning progress can be utilized to evaluate course characteristics, assess student performance, and forecast learning progress. Learning analytics, according to Lu et al. (2018), saves time for educators, which may be utilized to refine their teaching expertise and detect at-risk pupils at an earlier stage in the learning process. MOOCs, on the other hand, have basic flaws. Learning guidelines are lacking, and the completion rate is quite low (Freitas, Morgan, & Gibson, 2015). Low completion rates might be attributed to a variety of different circumstances. However, according to surveys, the vast majority of MOOC learners are unprepared for the enormous course content and the isolated learning environment that they would encounter (Kim, Olfman, Ryan, & Eryilmaz, 2014).

Data science is the interdisciplinary study of strategies and procedures for extracting information or insights from various data types, whether structured or unstructured (Irizarry, 2019). The phrase data science has recently developed to specifically denote a new profession expected to make sense of massive quantities of massive data. Educational data science (EDS) is the application of data science (DS) to educational data. Its roots can be traced back to computer science techniques such as computational statistics, data mining, machine learning, natural language processing, and human-computer interaction (Sweta, 2021).

The study of the above documents shows that MOOCs, big data, and data science are critical to students’ daily lives, especially in learning in the next normal. Including lifelong learning and online learning is an essential factor in learning via MOOCs. Therefore, the researcher has proposed an idea to design the system architecture of big data in Massive Open Online Courses (BD-MOOCs System Architecture) to be used as the guidelines for building information systems that teachers may be able to better guide their students and increase their motivation to learn.

2. Research Objectives

2) To design a system architecture of big data in massive open online courses (BD-MOOCs system architecture).

2) To evaluate the BD-MOOCs system architecture.

3. Literature Review

3.1 Analyzing Data and Improving Learning Effectiveness

Artificial intelligence (AI) is the process of simulating human intellect in robots so that their judgments and conclusions resemble those of a human mind (Akerkar, 2014; Su, Ding, & Chen, 2021; Su, Suen, & Hung, 2021). With the rising sophistication of data analytics in recent years, research on artificial intelligence in education (AIED) has developed (Kay & Kummerfeld, 2019; Schwendimann, 2017; Su & Lai, 2021; Su & Wu, 2021). Scholars obtain learner data from online platforms to examine learning processes (Daghestani, Ibrahim, AI-Towirgi, & Salman, 2020; Alexandron, Ruipérez-Valiente, Chen, MuozMerino, & Pritchard, 2017).

The rise of data analytics in education, particularly big data analysis, has cleared the way for a new teaching paradigm in which students’ actions and progress are observed in order to improve learning results. Students can also keep track of their progress in order to improve their self-directed learning (Kavitha & Raj, 2017). Hwang et al. (2020) created a fuzzy expert system-based adaptive learning strategy that took both affective and cognitive elements into account. The findings of the study suggested that the learning system could improve students' academic performance while also lowering their anxiety about learning.

Learning data analytics advances have resulted in the construction of a welcoming online learning environment that aids students in achieving their learning objectives, particularly in higher education distance teaching and teacher training courses. Teachers can track students' learning practices and evaluate their learning efficacy across multiple aspects using such devices (Meier, Xu, Atan, & Schaar, 2016).

3.2 Student Behavior Is Used to Evaluate Learning Performance

Learning behaviors are acts those students learn and employ to assess their learning and performance. Examining students' learning behaviors not only provides teachers with information about their students' learning situations, but it also guarantees that teaching materials are feasible. Hsu et al. (2021) created an instructional tool for AI education that recorded learning activities using videos and screenshots. When students were taught how to use AI, they discovered interesting behavioral patterns.

In learning assessments, students' MOOC learning behaviors are also a significant influence. MOOCs, on the other hand, frequently record low completion rates and significant dropout rates (Sun, Ni, Zhao, Shen, & Wang, 2019). Several research have offered ways for predicting students' course success or failure (Er et al., 2019; Lu et al., 2018). A logistic regression model is one such method for prediction. Lee (2018) used this method to investigate data from students' learning logs and study the behavior of students engaged in uninterrupted video watching. If students did not watch the course video for two days in a row, their learning was disrupted. From the learning log data, the author calculated the frequency and duration of uninterrupted learning actions and fed the information into the prediction model. Lee then established three continuous learning thresholds (10, 30, and 60 minutes) and evaluated the effects of continuous learning across the three thresholds. The precision-recall curve showed that the 60-minute threshold occupied the most area, indicating that it was the most beneficial in predicting student performance in acquiring a course certificate. In other words, students who participated in more learning activities and engaged in learning for a longer period of time are more likely to receive a course certificate.

Sun et al. (2019) proposed constructing a dropout prediction model using a gated recurrent unit-recurrent neural network (GRU-RNN) model. An RNN with a URL embedding layer forms the basis of the model. The authors utilized their model to compare student performance before and after starting the course, as well as to calculate the number of days students spent not learning. They then looked at other learning methods, such as completing exercise questions, participating in forums, and taking exams. Finally, in order to predict learning performance, the authors looked at students' learning patterns through a sequence of learning behaviors.

Pratsri et al. (2020) advocated retrieving sources from internal or external organizations and providing a platform for automating full operations in data collecting, storage, and analysis. Large data's tool is a big data issue. A big data platform for higher education must also consider security and privacy issues related to big data in corporations, government agencies, and educational institutions (HEi). It is a digital learning platform that enables online instruction and the use of digital media for educational reform, with a module that explains how computers and humans interact. A computer system that employs all of its computer resources to maximize efficiency (High-Performance Computing: HPC), as well as a network system to detect the target device network. After that, using Hadoop's tools and methodologies, the Big Data platform would perform necessary data analysis by accessing existing information. To show, accurate forecasting requires a lot of data about students and teachers.

Chinsook et al. (2022) proposed that big data is an important aspect of innovation that has recently piqued academic and practitioner interest. Given the importance of the education business, more research is being done on big data. Much research has been done to date to better comprehend big data's utilization in numerous areas for various reasons. The project aimed to (1) create the system architecture for big data in higher education for student behavior analytics and (2) evaluate the system architecture. The research process has two phases. The system and a five-level rating scale were used to collect data. Means and standard deviations were used to analyze the data. The Big Data-HE-SBA system architecture can be described as follows: a) Behavioral Analytics Big Data Sources; b) Behavioral Analytics Sub-Domains Big Data Sources; c) Behavioral Analytics Big Data Capture and Storage; d) Behavioral Analytics Big Data Behavioral AnalysisThe specialists' recommendations on the system architecture were excellent.

Pratsri et al. (2022) suggested a synthesis of data science performance for higher education students and a suitability assessment of data science performance for higher education students. The research instruments are as follows: 1) tables synthesizing data science performance, 2) expert interviews for data science performance assessments and 3) expert surveys to measure data science performance consistency. Analytical methods include the following: 1) analyzing the frequency of words in the content analysis table, 2) synthesis of content from interviews, and 3) analyzing the consistency and components of data science performance, derived from data science synthesis for higher education students, and concluding that data performance for higher education students is composed of five components: 1) programming skills, 2) basic statistics, 3) data science fundamentals, 4) data preparation, and 5) big data.

3.3 Measurement of Learner Skills in MOOCs

A judgment score or a standard reference is provided in traditional learning assessments. Students, on the other hand, have varying levels of learning capacity and quickness. Difficult exam questions do not accurately reflect the comprehension level of kids who have difficulty learning. To solve this problem, researchers developed test response theory, which has grown in popularity in both education research and practice. According to test response theory, students are given questions based on their responses to prior ones, and the test difficulty is thus adapted to the student's abilities. The idea, on the other hand, does not address how to correct student misconceptions or detect learning difficulties (Liu, Lin, & Tsai, 2009). A diagnosis can be done in a variety of ways. The most common qualitative method is interviews, and the most common quantitative method is test response theory. Diagnostic testing is an expanding subfield in the testing industry, because to the growing application of AI technologies, such as neural networks. Chu et al. (2020) sees cognitive diagnostic testing based on cognitive science theory as a major trend in the future. To see if cognitive science theory provides meaningful student ability evaluations, the author created a cognitive diagnostic test and presented a question-response methodology (Chu, Li, & Yu, 2020). Their approach improved learning data analytics, allowing MOOC instructors to better assess student performance and track learning habits across multiple learning dimensions.

The literature on MOOCs has extensively looked into online assessments and learner involvement. DeBoer, Ho, Stump, and Breslow (2014) investigated the idea of participation as well as desirable learning success and participation quality indicators. However, students may enroll in a course but fail to complete the examinations. The assessment quality of MOOCs was investigated by Admiraal, Huisman, and Van de Ven (2014). MOOCs involve a dynamic learning process in which students engage in a sequence of activities that include perception, learning, thinking, and problem-solving. As a result, final grades are an insufficient indicator of student achievement (Shepard, 2001). Because learning is a process rather than an outcome, teachers must watch students' learning habits throughout the course. The aforementioned conclusions highlight the necessity for MOOCs to use alternate assessment methodologies.

3.4 Exercises for Predicting Learning Performance

Moreno-Marcos, Pong, Muoz-Merino and Delgado-Kloos (2020) proposed a strategy for predicting students' assignment, examination, and final grades based on their learning status, performance in discussion forums, video-watching behaviors, practice question responses, and previous assignment scores. Previous assignment scores and average answer scores were found to be highly predictive of the three grades indicated above, however student performance in discussion forums was just somewhat predictive. Because some courses only provide videos with no exercises, the authors incorporated student behavioral data such as click counts as a model feature, but no significant improvement in performance was observed.

Huang, Chen, Tzeng and Lee (2018) used deep learning to create a concept assessment system using a knowledge map. Each week's knowledge topology was provided as a knowledge map by the writers. They gathered data on exercise difficulty and student behavior while watching videos and utilized it to predict students' comprehension of information in a given week's course. A deep learning strategy was used to create the prediction model.

A model to predict test scores was proposed by Li, Xie and Wang (2016). Using 15 features such as student age, gender, education level, registration time, number of videos watched, number of exercises completed, and related behaviors, the authors projected quiz marks based on numerous educational theories. The characteristics, however, were not significantly related to exam scores and so could not be employed in the prediction model.

3.5 MOOCs Lack Evaluation Methods

Standardized tests have typically been used to evaluate student achievement, thus learning tools that evaluate learning investments in hybrid, remote, or virtual learning contexts are needed. MOOCs have changed worldwide learning trends, however they confront numerous obstacles in terms of long-term development and learning models, such as poor completion rates (5–10%) and high learning loss rates (Sun, Ni, Zhao, Shen, & Wang, 2019). Because students cannot be observed in real time, evaluating learner success in MOOCs is intrinsically challenging, restricting MOOCs' ability to be unbiased or give accurate verification of coursework (Bady, 2013). Furthermore, MOOCs have a large number of students, and teachers are unable to communicate with each one individually. Students must rely on active interactions with other online learners to get learning feedback and practice in these situations. Students must also be self-directed learners (Crosslin, 2018). MOOC educators are working to develop online metrics for large-scale data collection for students of various levels and ages. Previous evaluation methods for online learners can serve as a guide for educators; however, MOOC educators are working to develop online metrics for large-scale data collection for students of various levels and ages.

4. Methodology

The purpose of the research is to design and evaluate a system architecture for big data in massive open online courses (BD-MOOCs system architecture). As follows, there are two steps in the research methodology that correspond to the research objectives:

4.1 Step 1

Design the system architecture for the BD-MOOCs.Data is gathered from papers and academic publications that include ideas, theories, and studies that are linked to the study architecture of big data for MOOCs. The information gathered via this technique emphasizes information that cannot be gained through an interview. For massive open online courses, it is necessary to design a BD-MOOCs system architecture.

4.2 Step 2

Population: The population is an expert in Big Data, Data Science, and Massive Open Online Courses. groups: The samples are ten experts in the fields of big data, data science, and Massive Open Online Courses. Chosen by purposive sampling. They are highly experienced experts in these fields, having worked for at least five years. Variable An independent variable is the BD-MOOCs system architecture. The dependent variable is the appropriateness of the BD-MOOCs system architecture. The research instruments were an evaluation of the system architecture of big data science in education for Massive Open Online Courses. The data obtained from the experts was collected and analyzed by using the statistics as follows: The arithmetic mean and standard deviation are computed. Ten experts in the evaluation of the system architecture of big data science in education for Massive Open Online Courses. The instrument used is the suitability assessment form for the BD-MOOCs system architecture. They're broken down into five levels, with the following criteria for evaluating and interpreting each one:

Table 1. Mean score range and interpretation of results

Average score range	Meaning of interpretation
4.50–5.00	Level 5 means the suitability at the highest level
3.50–4.49	Level 4 means the suitability at the high level
2.50–3.49	Level 3 means the suitability at the medium level
1.50–2.49	Level 2 means the suitability at the low level
1.00–1.49	Level 1 means the suitability at the lowest level

5. Results

5.1 A System Architecture of Big Data in Massive Open Online Courses (BD-MOOCs System Architecture)

This diagram indicates all of the most important components of the BD-MOOCs System Architecture.

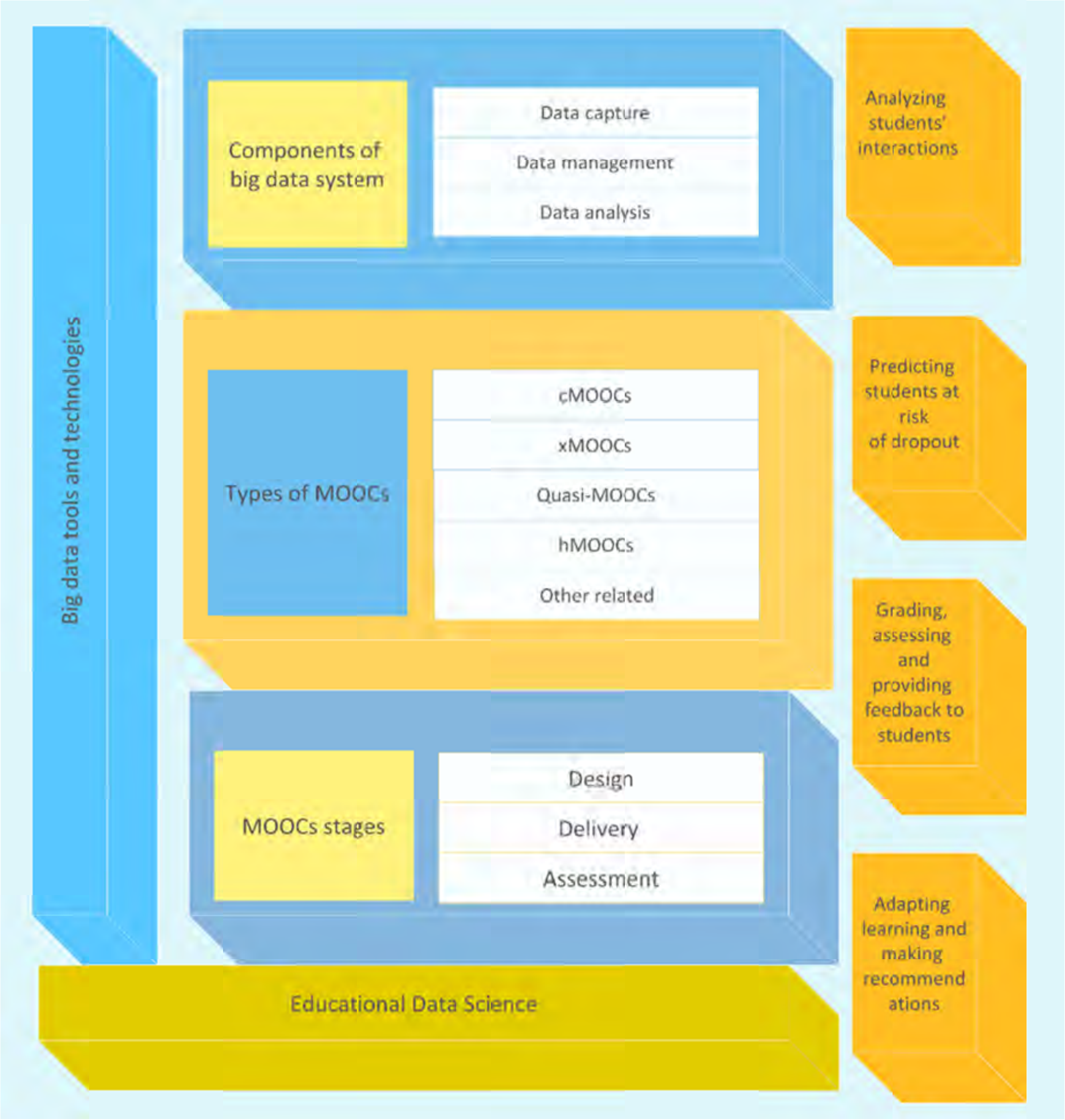
Big data tools and technologies
Components of big data system
Data capture
Data management
Data analysis
Analyzing students' interactions
Types of MOOCs
cMOOCs
xMOOCs
Quasi-MOOCs
hMOOCs
Other related
Predicting students at risk of dropout
Grading, assessing and providing feedback to students
MOOCs stages
Design
Delivery
Assessment
Adapting learning and making recommend ations
Educational Data Science

Figure 1. System architecture of big data in massive open online courses (BD-MOOCs System Architecture)

1) Big data tools and technologies

Many novel data storage and processing methods were launched during the big data era. Hadoop, MapReduce, Spark, NoSQL databases, in-memory computing, distributed file storage, and predictive analytics are just a handful of the numerous technologies available for constructing big data systems. We offer just a few of the technologies employed in developing our system architecture that integrates e-learning systems into the big data ecosystem.

- Apache Hadoop is an open-source framework for storing and analyzing large amounts of data. This technology permits the processing and distribution of enormous data volumes simultaneously. It is meant for hardware usage. It is offered in three modes: independently, pseudo-distributed, and wholly distributed. The Hadoop software library consists in the practice of many related projects, including storage, querying (Pig), transferring and streaming, and providing associated services. The Hadoop Dispersed File System (HDFS) stores information, and MapReduce is used to analyze large volumes of data scattered over a cluster of computers. It consists of two key elements.

- YARN (Yet Another Resource Negotiator) is an Apache Hadoop framework for task scheduling and cluster resource management. It is intended to support several data computing frameworks such as Spark, Storm, and others. It features a simple and adaptable design based on the master/worker idea, with the master referred to as Resource Manager (RM) and the workers referred to as Node Managers (NM). The RM is in charge of monitoring and allocating resources among a cluster of devices. On the other hand, NM is in control of job execution and resource use per worker node.
- The Hadoop Distributed File System is a file system for data storage and administration. It is intended for use on commodity hardware. It is also fault-tolerant and allows high-throughput data access. It is also appropriate for applications with enormous datasets. The two primary components of HDFS are Name Node, which maintains the naming system (block locations) of the files or directories, and Data Nodes, which are used for storing, reading, duplicating, and deleting a block of data stored on distributed nodes.
- Apache Spark is a very efficient data processing engine. It is a distributed computing platform that provides excellent performance for both batch and interactive processing. Because it conducts in-memory data processing, Spark enables applications on Hadoop clusters to operate up to 100 times quicker than Hadoop MapReduce, decreasing the actual number of reading and write operations. As a result, there is no time wasted transferring data in and out of the disk. High-level APIs are available in Java, Scala, Python, and R. It also supports a wide range of higher-level tools such as Spark SQL, Spark Streaming, MLlib, and GraphX. Spark functions in several environments such as Hadoop, Mesos, independent cloud computing, and others. It may link to several sources such as HDFS, MongoDB, and HBase.
- Apache Hive is a data warehouse that efficiently queries and manages massive amounts of data stored in distributed storage. It features HiveQL, a SQL-like query language, for data management. In addition, a tool for the command line and the JDBC driver is provided to connect Hive users, which enables several actions such as extracting/turning/loading. It also provides fast access to files stored on different data storage systems such as HDFS and HBase.
- Apache Sqoop is a framework for moving data across organized, semi-structured, and unstructured data sources efficiently. It may import and export data to HDFS or other data-storage systems like HBase and Cassandra from structured data sources, such as related databases systems.
- Apache Flume is an abbreviation for a distributed, dependable, and accessible data storage system for gathering, storing, and transporting massive volumes of log data from several sources. Furthermore, a fundamental, extendable data architecture enables online analytical applications. Flume’s architecture is basic. It consists of three parts: a flume agent, a source, a channel, and a sink. Indeed, flume source takes data flow units from an external data source known as events (log file). The events received are then stored on the flume canal.

2) Educational data science

- Educational Data Mining (EDM) examines activities at the micro or “understanding” level to determine which teaching approaches and academic interventions are most likely to improve learners’ learning of a specific topic.
- Educational Recommender Systems (ERS) investigate how variations among learners impact their persistence and how overall institute performance may be boosted by utilizing adaptation, personalization, and recommendation strategies at the macro or “learner” level.
- Academic/Institutional Analytics (AA), formerly known as Institutional Research, focuses on the institution rather than learning processes or specifics about where learning occurs.
- Systemic/Instructional Improvement (S/II), also known as Data-Driven Decision Making, evaluates teaching approaches using data from test-based systems and state longitudinal data systems.

3) Components of the big data system

- Data capture: The major duties are data creation and data acquisition. Data may be collected from a variety of sources, including server logs, sensors, and search engines. The vast majority of data sources are user-generated and machine-generated data, which combine to build large databases.
- Data management: The next crucial stage is data management, separated into two sub-tasks when data are produced from various sources. The terminology is used to describe data storage and recovery. Big data applications work on a huge scale. The distributed filesystem handles large-scale data storage. The second task that NoSQL databases deal with is data recuperation. MapReduce is also a distributed computing architecture that enables massive data sets to be dispersed across a node network. We illustrate the characteristics of a distributed file system and a comparison of several file systems based on several factors.
- Data analysis: Data Analysis is the last step in extensive data processing. The two most essential subtasks of data analysis are data analytics and data visualization. Data is modeled using ML algorithms and analyzed using statistical/data mining approaches to uncover hidden patterns in analytics.

4) Types of MOOCs

- cMOOCs or connectivism MOOCs, as depicted by early-era ideals, emphasize massively dispersed peer learning and are part of the open education resource movement.
- xMOOCs, which are unidirectional and often based around traditional lecture forms, is now the most common kind. They are increasingly being given through proprietary learning management systems like Coursera, Udacity, edX, MITx, and others.
- Quasi-MOOCs include a wide range of web-based tutorials, such as OER (Open Educational Resources), which are not officially courses but are meant to facilitate learning-specific activities.
- hMOOCs, also known as hybrid MOOCs or MOOCs 3.0, enable hybrid or flipped classrooms that blend online and in-person learning experiences.
- Other related terms used for this kind, of course, include BOOC (Big Open Online Course), aMOOC (adaptive MOOC), bMOOC (blended MOOC), sMOOC (semi-massive open online course), and so on.

5) MOOCs stages

- Design: Agents might do data interpretation tasks that are not accessible in typical MOOC apps by using their talents. By monitoring the MOOC environment and other associated system data, agents might use information such as use patterns, navigation, content sequencing, problematic content sections or data formats, tool use, and student profile. Course creators might then use this data to optimize MOOC adoption for future cohorts.
- Delivery: Content customization: Modern learners want to maximize the time they spend on learning activities and their efficacy in terms of their unique talents, knowledge, preferences, and learning goals. They would like the MOOC environment to recognize their learning style and adapt the learning situation accordingly in terms of specific content, didactic approaches, the type of media to be used, the order in which concepts to be learned are sequenced, and so on, while also receiving adaptive/personalized feedback to improve their performance and motivation.
- Assessment: Agents might be employed to enhance present automated tests by modifying assignment questions according to the participant’s educational level or modifying the sequencing of the assessment questions if participants fail or pass a question before that one. Furthermore, by incorporating assessment approaches such as student e-portfolios, learning analytics, and enhanced rubrics, agents might aid in the development of a rigorous automated assessment system (including self-assessment rubrics and peer assessment rubrics).

6) Challenges in educational data science posed by MOOCs

- Analysing students’ interactions: Analysis of students’ MOOC interaction data may give lecturers, resource developers, and members of organizations helpful information to enhance their MOOCs by highlighting critical course difficulties. A typical MOOC includes various data sources related to student activities, such as use and engagement, video lectures, and social networking communication through forums. Every click, page, slide read or viewed contribution, video player command, test or question answered, and social engagement on a forum may all leave a digital trace. However, manually analyzing the massive quantity of student interaction data collected across several MOOCs is tough. As a result, EDS procedures must be used.

- Predicting students at risk of dropout: Compared to typical university courses, MOOCs have lower completion rates, and this high student dropout rate has been used to throw doubt on its promised potential. The funnel of participation metaphor depicts the rapid drop-off inactivity and the pattern of severely uneven involvement that seems to be distinctive of MOOCs. There are several plausible explanations for poor MOOC retention rates, including a lack of time, learners' motivation, a sense of isolation, a lack of involvement, inadequate prior knowledge and abilities, and hidden expenses. Several EDS approaches (mostly correlation, regression, and classification) have been used to forecast students at risk of dropping out.
- Grading, assessing, and providing feedback to students: The massiveness and a lack of instructional participation are two of the most severe issues with MOOCs. Because of the massive participation in MOOCs, which may result in a student-teacher ratio of ten thousand to one or more, the time an instructor spends analyzing each student's work and offering comments is minimal. This absence of connection between MOOC instructors and participants may irreparably harm the course's quality.
- Adapting learning and making recommendations: The heterogeneity of MOOC learners is excellent because of the vast number of concurrent participants from various educational and cultural backgrounds and varying levels of maturity and experience. Participants may experience dissatisfaction due to a lack of flexibility to their unique requirements and learning styles. One solution is to employ adaptive learning systems, which create a feeling of customization based on a particular learner model, and recommender systems, which deliver suggestions to an individual student or a group of students.

5.2 The Evaluation of of the BD-MOOCs Model

Table 2. The assessment of big data tools and technologies

Evaluation Lists	***Level of suitability***	**Level of assessment**	
		$\bar{x}$	**S.D.**
1. Apache Hadoop	highest	4.70	0.483
2. YARN	highest	4.60	0.516
3. Hadoop Distributed File System	highest	4.70	0.483
4. Apache Spark	highest	4.80	0.422
5. Apache Hive	highest	4.50	0.527
6. Apache Sqoop	highest	4.60	0.516
7. Apache Flume	highest	4.60	0.516
Overall	**highest**	**4.64**	**0.495**

The following Table 2 describes the assessment of big data tools and technologies by the experts, with suitability at the highest level ($\bar{x}$ = 4.64, S.D. = 0.495).

Table 3. The educational data science assessment

Evaluation Lists	***Level of suitability***	**Level of assessment**	
		$\bar{x}$	**S.D.**
1. Educational Data Mining (EDM)	highest	4.70	0.483
2. Recommender Systems (ERS)	highest	4.50	0.527
3. Academic/Institutional Analytics (AA)	highest	4.50	0.527
4. Systemic/Instructional Improvement (S/II)	highest	4.80	0.422
Overall	**highest**	**4.62**	**0.489**

The following Table 3 describes the educational data science assessment by the experts, with suitability at the highest level ($\bar{x}$ = 4.62, S.D. = 0.489).

Table 4. The assessment of components of big data systems

Evaluation Lists	*Level of suitability*	Level of assessment	
		x̅	S.D.
1. Data capture	highest	4.90	0.316
2. Data management	highest	4.50	0.527
3. Data analysis	highest	4.70	0.483
Overall	**highest**	**4.70**	**0.442**

The following Table 4 describes the assessment of components of big data systems by the experts, with suitability at the highest level (x̅ = 4.70, S.D. = 0.442).

Table 5. The assessment of types of MOOCs

Evaluation Lists	*Level of suitability*	Level of assessment	
		x̅	S.D.
1. cMOOCs	highest	4.50	0.527
2. xMOOCs	highest	4.50	0.527
3. Quasi-MOOCs	highest	4.90	0.316
4. hMOOCs	highest	4.50	0.527
5. Other related	highest	4.80	0.422
Overall	**highest**	**4.64**	**0.463**

The following Table 5 describes the assessment of types of MOOCs by the experts, with suitability at the highest level (x̅ = 4.64, S.D. = 0.463).

Table 6. The assessment of MOOC stages

Evaluation Lists	*Level of suitability*	Level of assessment	
		x̅	S.D.
1. Design	highest	4.60	0.516
2. Delivery	highest	4.60	0.516
3. Assessment	highest	4.70	0.483
Overall	**highest**	**4.63**	**0.505**

The following Table 6 describes the assessment of MOOC stages by the experts, with suitability at the highest level (x̅ = 4.63, S.D. = 0.505).

Table 7. The assessment of Challenges in educational data science posed by MOOCs

Evaluation Lists	*Level of suitability*	Level of assessment	
		x̅	S.D.
1. Analysing students' interactions	highest	4.60	0.516
2. Predicting students at risk of dropout	highest	4.60	0.516
3. Grading, assessing and providing feedback to students	highest	4.70	0.483
4. Adapting learning and making recommendations	highest	4.60	0.516
Overall	**highest**	**4.62**	**0.507**

The following Table 7 describes the assessment of challenges in educational data science posed by MOOCs by the experts, with suitability at the highest level (x̅ = 4.62, S.D. = 0.507).

Table 8. Summarizes the BD-MOOCs system architecture results.

Evaluation Lists	***Level of suitability***	**Level of assessment**	
		$\bar{x}$	**S.D.**
1. Big data tools and technologies	highest	4.64	0.495
2. Educational data science	highest	4.62	0.489
3. Big data systems	highest	4.70	0.442
4. Types of MOOCs	highest	4.64	0.463
5. MOOC stages	highest	4.63	0.505
6. Challenges in educational data science posed by MOOCs	highest	4.62	0.507
Overall	**highest**	**4.64**	**0.483**

The following Table 8 summarizes the BD-MOOCs system architecture results by MOOCs, with suitability at the highest level ($\bar{x}$ = 4.64, S.D. = 0.483).

6. Conclusion

Big data in education paves the way for the future creation of MOOCs that actually provide true learning freedom. Simultaneously, we would want to advise any open universities or institutions interested in promoting their own connotation and brand building through MOOC creation to consider the influence of the big data education paradigm while constructing an effective online learning environment. This will enable more informed MOOC design decisions and programs that are more inspiring and pedagogically effective. According to the system architecture of big data in massive open online courses (BD-MOOCs System Architecture), the following findings were drawn about the study's system architecture. The BD-MOOCs system architecture, which comprises six components, was developed to generate a vast amount of big data in massive open online courses. The first component was comprised of big data tools and technologies, including Apache Hadoop, YARN, the Hadoop Distributed File System, Apache Spark, Apache Hive, Apache Sqoop, and Apache Flume. The second component is educational data science, which is divided into four parts: EDM, ERS, AA, and S/II. The third feature was that big data systems were divided into three elements: data capture, data management, and data analysis. The fourth part was the classification of MOOCs into five categories: cMOOCs, xMOOCs, quasi-MOOCs, hMOOCs, and other related. The fifth component was that the MOOC phases consisted of three components: design, delivery, and assessment. The final component was the four aspects of the educational data science challenges posed by MOOCs: analyzing student interactions; predicting students at risk of dropping out; grading; and adapting learning and making recommendations. In this case, the assessment of the BD-MOOCs system architecture design was suitable at the highest level.

References

Admiraal, W., Huisman, B., & Van de Ven, M. (2014). Self-and peer assessment in massive open online courses. *International Journal of Higher Education*, *3*(3), 119–128. https://doi.org/10.5430/ijhe.v3n3p119

Akerkar, R. (2014). *Introduction to artificial intelligence*. Patparganj, India: PHI Learning Pvt. Ltd.

Alexandron, G., Ruiperez-Valiente, J. A., Chen, Z. Z., Munoz-Merino, P. J., & Pritchard, D. E. (2017). Copying@Scale: Using harvesting accounts for collecting correct answers in a MOOCs. *Computers & Education*, *108*, 96–114. https://doi.org/10.1016/j.compedu.2017.01.015

Ashabi, A., Sahibuddin, S. B., & Haghighi, M. S. (2020). *Big Data: Current Challenges and Future Scope*. 2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE). https://doi.org/10.1109/ISCAIE47305.2020.9108826

Bady, A. (2013). The MOOCs moment and the end of reform. *Liberal Education*, *99*(4), 6–15.

Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. *Mobile Networks and Applications*, *19*(2), 171–209. https://doi.org/10.1007/s11036-013-0489-0

Chinsook, K., Khajonmote, W., Klintawon, S., Sakulthai, C., Leamsakul, W., & Jantakoon, T. (2022). Big Data in Higher Education for Student Behavior Analytics (Big Data-HE-SBA System Architecture). *Higher Education Studies*, *12*(1), 105–114. https://doi.org/10.5539/hes.v12n1p105

Chu, L., Li, P. H., & Yu, M. N. (2020). The Longitudinal effect of children's self-regulated learning on reading habits and well-being. *International Journal of Educational Research*, *104*, 101673. https://doi.org/10.1016/j.ijer.2020.101673

Crosslin, M. (2018). Exploring self-regulated learning choices in a customizable learning pathway MOOCs.

Australasian Journal of Educational Technology, *34*(1), 131–144. https://doi.org/10.14742/ajet.3758

Daghestani, L. F., Ibrahim, L. F., Al‐Towirgi, R. S., & Salman, H. A. (2020). Adapting gamified learning systems using educational data mining techniques. *Computer Applications in Engineering Education*, *28*(3), 568–589. https://doi.org/10.1002/cae.22227

DeBoer, J., Ho, A. D., Stump, G. S., & Breslow, L. (2014). Changing "course" reconceptualizing educational variables for massive open online courses. *Educational Researcher*, *43*(2), 74–84. https://doi.org/10.3102/0013189X14523038

Er, E., Gomez-Sanchez, E., Dimitriadis, Y., Bote-Lorenzo, M. L., Asensio-Perez, J. I., & Alvarez-Alvarez, S. (2019). Aligning learning design and learning analytics through instructor involvement: A MOOCs case study. *Interactive Learning Environments*, *27*(5–6), 685–698. https://doi.org/10.1080/10494820.2019.1610455

Freitas, S. I., Morgan, J., & Gibson, D. (2015). Will MOOCs transform learning and teaching in higher education? Engagement and course retention in online learning provision. *British Journal of Educational Technology*, *46*(3), 455–471. https://doi.org/10.1111/bjet.12268

Hsu, T. C., Abelson, H., Lao, N., Tseng, Y. H., & Lin, Y. T. (2021). Behavioral-pattern exploration and development of an instructional tool for young children to learn AI. *Computers and Education: Artificial Intelligence*, *2*, 1–14. https://doi.org/10.1016/j.caeai.2021.100012

Huang, N. F, Chen, C. C, Tzeng, J. W., & Lee, C. A. (2018). *Concept assessment system integrated with knowledge map using deep learning* (pp. 113–116). Proceedings of the IEEE Learning with MOOCs (LWMOOCs 2018). Madrid, Spain. https://doi.org/10.1109/LWMOOCS.2018.8534674

Hwang, G. J., Chu, H. C., & Yin, C. (2017). Objectives, methodologies and research issues of learning analytics. *Interactive Learning Environments*, *25*(2), 143–146. https://doi.org/10.1080/10494820.2017.1287338

Hwang, G. J., Sung, H. Y., Chang, S. C., & Huang, X. C. (2020). A Fuzzy expert system-based adaptive learning approach to improving students' learning performances by considering affective and cognitive factors. *Computers and Education: Artificial Intelligence*, *1*, 1–15. https://doi.org/10.1016/j.caeai.2020.100003

Irizarry, R. A. (2019). *Introduction to Data Science*. https://doi.org/10.1201/9780429341830

Islam, M., & Reza, S. (2019). The Rise of Big Data and Cloud Computing. *Internet of Things and Cloud Computing*, *7*(2), 45. https://doi.org/10.11648/j.iotcc.20190702.12

Kavitha, G., & Raj, L. (2017). Educational data mining and learning analytics - Educational assistance for teaching and learning. *International Journal of Computer & Organization Trends*, *41*(1), 21–25. https://doi.org/10.14445/22492593/IJCOT-V41P304

Kay, J., & Kummerfeld, B. (2019). From data to personal user models for life-long, life-wide learners. *British Journal of Educational Technology*, *50*(6), 2871–2884. https://doi.org/10.1111/bjet.12878

Kim, R., Olfman, L., Ryan, T., & Eryilmaz, E. (2014). Leveraging a personalized system to improve self-directed learning in online educational environments. *Computers & Education*, *70*, 150–160. https://doi.org/10.1016/j.compedu.2013.08.006

Kinskey, C., King, H., & Lewis Miller, C. (2018). Open educational resources: An analysis of Minnesota State Colleges and Universities student preferences. *Open Learning: The Journal of Open, Distance and E-Learning*, *33*(3), 190–202. https://doi.org/10.1080/02680513.2018.1500887

Kler, S. (2014). ICT Integration in Teaching and Learning: Empowerment of Education with Technology. *Issues and Ideas in Education*, *2*(2), 255–271. https://doi.org/10.15415/iie.2014.22019

Kockum, F., & Dacre, N. (2021). *Project Management Volume, Velocity, Variety: A Big Data Dynamics Approach*. https://doi.org/10.31235/osf.io/k3h9r

Lee, Y. (2018). Effect of uninterrupted Time-On-Task on students' success in massive open online courses (MOOCs). *Computers in Human Behavior*, *86*, 174–180. https://doi.org/10.1016/j.chb.2018.04.043

Li, X., Xie., L., & Wang, H. (2016). *Grade prediction in MOOC*. Proceedings of the 2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES), Paris, France. https://doi.org/10.1109/CSE-EUC-DCABES.2016.213

Liu, T. C., Lin, Y. C., & Tsai, C. C. (2009). Identifying misconceptions about statistical correlation and their possible causes among high school students: An Exploratory study using concept mapping with interviews. *International Journal of Science and Mathematics Education*, *7*(4), 791–820. https://doi.org/10.1007/s10763-008-9142-y

Lu, O., Huang, A., Huang, J., Lin, A., Ogata, H., & Yang, S. (2018). Applying learning analytics for the early prediction of students' academic performance in blended learning. *Educational Technology & Society*, *21*(2), 220–232.

Meier, Y., Xu, J., Atan, O., & Schaar, M. V. D. (2016). Predicting grades. *IEEE Transactions on Signal Processing*, *64*(4), 959–972. https://doi.org/10.1109/TSP.2015.2496278

Moreno-Marcos, P. M., Pong, T., Muñoz-Merino, P. J., & Delgado Kloos, C. (2020). Analysis of the factors influencing learners' performance prediction with learning analytics. *IEEE Access*, *8*, 5264–5282. https://doi.org/10.1109/ACCESS.2019.2963503

Pratsri, S., & Nilsook, P. (2020). Design on Big data Platform-based in Higher Education Institute. *Higher Education Studies*, *10*(4), 36–43. https://doi.org/10.5539/hes.v10n4p36

Pratsri, S., Nilsook, P., & Wannapiroon, P. (2022). Synthesis of Data Science Competency for Higher Education Students. *International Journal of Education and Information Technologies*, *16*, 101–109. https://doi.org/10.46300/9109.2022.16.11

Rao, T. R., Mitra, P., Bhatt, R., & Goswami, A. (2018). The big data system, components, tools, and technologies: A survey. *Knowledge and Information Systems*, *60*(3), 1165–1245. https://doi.org/10.1007/s10115-018-1248-0

Romero, C., & Ventura, S. (2016). Educational data science in massive open online courses. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(1). https://doi.org/10.1002/widm.1187

Saadia, D. (2021). Integration of Cloud Computing, Big Data, Artificial Intelligence, and Internet of Things. *International Journal of Web-Based Learning and Teaching Technologies*, *16*(1), 10–17. https://doi.org/10.4018/IJWLTT.2021010102

Schwendimann, B. A. (2017). Perceiving learning at a glance: A Systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies*, *10*(1), 30–41. https://doi.org/10.1109/TLT.2016.2599522

Shepard, L. (2001). *The Role of classroom assessment in teaching and learning*. CSE Technical Report. Los Angeles, USA: National Center for Research on Evaluation, Standards, and Student Testing.

Su, Y. S., Ding, T. J., & Chen, M. Y. (2021). Deep learning methods in internet of medical things for valvular heart disease screening system. *IEEE Internet of Things Journal*. https://doi.org/10.1109/JIOT.2021.3053420

Su, Y. S., & Lai, C. F. (2021). Applying educational data mining to explore viewing behaviors and performance with flipped classrooms on the social media platform Facebook. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.653018

Su, Y. S., Suen, H. Y., & Hung, K. E. (2021). Predicting behavioral competencies automatically from facial expressions in real-time video recorded interviews. *Journal of Real-Time Image Processing*. https://doi.org/10.1007/s11554-021-01071-5

Su, Y. S., & Wu, S. Y. (2021). Applying data mining techniques to explore users behaviors and viewing video patterns in converged IT environments. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-020-02712-6

Sun, D., Mao, Y., Du, J., Xu, P., Zheng, Q., & Sun, H. (2019). *Deep learning for dropout prediction in MOOCs* (pp. 87–90). Proceedings of the 2019 Eighth International Conference on Educational Innovation through Technology (EITT). Mississippi, USA. https://doi.org/10.1109/EITT.2019.00025

Sun, Y. Q., Ni, L. H., Zhao, Y. M., Shen, X. L., & Wang, N. (2019). Understanding students' engagement in MOOC: An integration of self-determination theory and theory of relationship quality. *British Journal of Educational Technology*, *50*(6), 3156–3174. https://doi.org/10.1111/bjet.12724

Sweta, S. (2021). Educational Data Mining Techniques with Modern Approach. *Modern Approach to Educational Data Mining and Its Applications*, 25–38. https://doi.org/10.1007/978-981-33-4681-9_3

Toth, C. T. (2019). *Massive Open Online Courses* (MOOCs) (pp. 29–106). Massive Open Online Courses Im

Kontext Von Persönlichkeit Und Prokrastination. https://doi.org/10.1007/978-3-658-26296-9_3

UNESCO. (2002). *Forum on the impact of open courseware for higher education in developingcountries.* Final report. Retrieved from http://www.unesco.org/iiep/eng/focus/opensrc/PDF/OERForumFinalReport.pdf

Copyrights