Received: 21 June 2016 | Accepted: 8 May 2017
DOI: 10.1002/cae.21844

REVIEW

WILEY

Data science in education: Big data and learning analytics

Aleksandra Klašnja-Milićević | Mirjana Ivanović | Zoran Budimac

Faculty of Sciences, Department of Mathematics and Informatics, University of Novi Sad, Novi Sad, Serbia

Correspondence
Aleksandra Klašnja-Milićević, Faculty of Sciences, Department of Mathematics and Informatics, Trg Dositeja Obradovića 3, University of Novi Sad, Novi Sad, Serbia.
Email: akm@dmi.uns.ac.rs

Abstract

This paper considers the data science and the summaries significance of Big Data and Learning Analytics in education. The widespread platform of making high-quality benefits that could be achieved by exhausting big data techniques in the field of education is considered. One principal architecture framework to support education research is proposed.

KEYWORDS
big data, education, learning analytics, learning environments

1 | INTRODUCTION

The Internet and recently also mobile computing and cloud computing completely changed our perception and attitude toward data. Data can be collected from more people and sources extremely cheap, for long time intervals, and with less effort than ever before [27,28]. This raised greater research challenges and opportunities and triggered off a wide range of innovative applications in many domains.

Big data recently appeared as a buzzword, or catch phrase, extensively has been used in inter and multi-disciplinary areas. It is believed that the term has originated with Web search companies which intention was to extract useful information from huge and spread collections of data with poorly structures [29]. Big data describe a large and great volume of data (structured or unstructured) difficult to be processed by traditional and standard software and database techniques. In contemporary business environments the data is huge, it moves extremely fast or it surpasses processing capability. Big data and appropriate processing techniques offer great potential for improving business in companies and provide support for more intelligent decisions. As a consequence of rapid innovations in technology, big data can nowadays be analyzed, interpreted, and providing new insights and benefits to areas like government, healthcare, e-learning, and in a wide range of data-driven industries [4,29].

Definition of big data be influenced by different factors like: the capacity, speediness, and cost of computing, and storage technologies. Big data characteristics rapidly change over the years. The amount of data that can be regarded as big data in the 1980s was 100 GB. Several decades later, by 2010, big data was 13 Petabyte. On the other hand, text file with 10 GB size can be also perceived as big data because usual text editors could not handle file with such size. Accordingly, the understanding and characterization of big data depends not only on a enormous amount of data however also have to consider which size of big data could be handled with selected technology [11].

During the last 20 years in various fields, data have increased enormously. International Data Corporation[1] (IDC) reported in 2011 that available data volume in the world was 1.8 ZB ($\approx 10^{21}$B). According to them data amount increases almost nine times within 5 years. As big data typically consists of unstructured data, the essential research challenge is how to effectively organize, analyze, and manage such datasets, especially in different popular applications: search engines, online shops, social media, and so on.

Researchers and data analysts, scientific and technological enterprises, and wide range of practitioners have been proposing different definitions of big data. One of the earliest definitions was proposed in 2001 by Laney [25]: "Big data can be characterized by the three Vs "volume" (expressed in terms of terabytes, records, transactions, tables, files),

[1] http://www.idc.com/

"velocity" (expressed in terms of batch, near time, real time, streams) and "variety" (expressed in terms of structured, unstructured, semi-structured, or of all mentioned)." Gartner[2] Inc. defines big data in similar terms: "Big data is high-volume, high-velocity, and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

Apache Hadoop (www.cloudera.com) in 2010 defined big data as "datasets which could not be captured, managed, and processed by general computers within an acceptable scope."

In 2011, an IDC report defined big data as "big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and/or analysis." According to this big data are summarized as four Vs: Volume (very large volume of data), Variety (various modalities), Velocity (rapid generation), and Value (huge value but very low density). After that, this 4 Vs definition has been widely accepted as it indicates a serious problem in big data: how to find out the values from datasets with a massive scale, diverse types, and hasty generation.

Closely related to the term "big data" is the term "big data analytics." It is the process of gathering, unifying analyzing, and evaluating huge sets of data with the intention to determine patterns of data and some other important information [4,5]. Big data analytics are an instrument for discovering the knowledge that is behind the analyzing data. Big data analytics have to use specialized high-performance analytics techniques and tools for data mining, data optimization, predictive analytics, forecasting, and so on [41].

In this paper, we will concentrate on big data, data/learning analytics, and their significance in education environments. The second section gives an overview of key trends in education environments. The third section is devoted to big data tools and techniques and their role in contemporary education. The challenges of implementation of big data techniques in education environments are pointed out in the section 4. The fifth section of the paper provides the concluding remarks and possible future directions relating to the development and implementation of big data education environments.

2 | KEY TRENDS IN CONTEMPORARY EDUCATIONAL ENVIRONMENTS

The Internet and mobile computing completely changed our perception of and relation to data. More people can collect data, with less effort, across long periods of time, at a lower cost than in past [38]. This offers more challenges and opportunities for a wide range of domains including education. We are witnesses of the existence of a critical mass of technologists and educators, who support activities and expect great promise in the analysis of such data. In contrast, there are concerns that the utilization of this data could bring different problems and obstacles to education and society in general. Nevertheless, full impact and expand of data sciences in education is expected to happen in near future.

Developments of new technologies such as digital learner records, learner cards, sensors, and mobile devices, flexible classroom design, and Massive Open Online Course (MOOC) is completely transforming the approach of learning and teaching [40]. Higher education institutions are collecting more data than ever before. Management of this vast amount of data should offer valuable comprehension about the learning process, insights about risk of learner's dropping out, and support for increasing learners' success. In order to comprehend the patterns of value that exist within the large amount of data new or innovative approaches are required. Lot of exploration and researches aim to handle the data with the proper techniques and new tools to produce real time solution, prediction in this certain area. The utilities of proper techniques and new tools could be: operative self-learning, useful peer groups, available class time for creativeness, and possibilities for problem solving [34].

Big data can address some of the key challenges in higher education practice:

1. improving learners' experience [42],
2. improving learners' knowledge trough enhanced academicstudying,
3. more effective evidence-based decision making,
4. strategic response to changing global trends [15],
5. opportunity for converting complex, often unstructured data into actionable information.

Each year The New Media Consortium [30] and EDUCAUSE publish The New Horizon Reports [23]. The main purpose of the reports is to cover research of educational community and present opportunities for modeling learning collaborations within large-scale data collection. In the last reports of higher education edition, it expects six evolving technologies over the next 5 years (Horizon report, 2012):

1. mobile applications,
2. computing on tablet,
3. learning analytics,
4. internet of things,
5. game-based learning,
6. gesture-based learning.

[2] http://www.gartner.com/technology/home.jsp

Big data providers, Coursera,[3] edX,[4] and Udacity,[5] serve several thousands of registered learners [13]. MOOC provides a plenty of opportunities:

1. advanced and continued learning,
2. low cost,
3. life-long learning,
4. new skills and improve knowledge of learners and teachers [35].

MOOCs can include thousands of learners and it is expected to grow in number and influence within the next years [34]. The MOOC[6] structure tends to be asynchronous and flexible to accommodate the varying levels of participation. Anyone can participate for free in any or the entire course's learning activities (e.g., blogs, video lectures, discussions, and other social media tools). The main shortcomings of MOOC are variability across and within courses and the lack of learner achievement rates. MOOCs are not suitable for all learners; especially those who like structured way of learning. Critics highlight that there is a need for detail inspecting these new approaches to ensure they are effective and advance past the traditional lecture-style pedagogies [37].

Nowadays, in higher education environments, it is evident that partnerships between companies and academia are demanding to achieve higher quality educational outcomes and such cooperation is constantly increasing [26]. Companies involve institutions of higher education to easily determine innovative technologies with the purpose of increasing abilities for research outputs, knowledge transfer, and commercialization [2]. Simultaneously, these continuous development of learning technologies and overall environment changes have impact on the higher education institutions, academic teaching, and researching [14]. In addition, many institutions need to implement new technologies aimed to meet growing learner needs and reduce operational costs.

Researchers of The U.S. Education Department emphasized that the "following 5 years will enhance the collaboration models between designers, investigators, and instructors" [6]. These models involve many participants in the analytics process: students, faculties, education institutions, scientists, government funders, and establishments. A complete view of learning analytics takes account of an extensive collection of learning activities, as well as the full learner experience: education before enrolment at the university, design of learning process, and evaluation of educational needs. To improvement learning analytics, researchers have to resolve the following features [6]:

1. enhancement of innovative techniques, tools and people;
2. accessibility, integrity, and range of data;
3. mark analytics activity; and
4. influences to correlated areas and experts.

Data mining technique and learning analytics methods can be very useful in education field. For example, Baker and Siemens [2] discussed four types of analysis that could bring new quality and significant benefits in learning activities.

1. *Prediction* on future uses of the learning environment—based on available learning data collected during learning processes and activities, it is possible to predict future uses of learning sequences, predict on final learners' grades, or predict students' knowledge behavior.
2. *Structure discovery*—based on available learning data it is possible to determine significant relations and patterns between knowledge levels of students, times of e-learning system usage, and students' grades.
3. *Relationship mining*—on the basis of the information gathered from the interaction between the user and the learning environment, it is possible to discover relationship between the usability of the course materials and the students' learning performances.
4. *Distillation of data*—also it is important to distill data in different ways and for different purposes for further use in human management.

For the prediction type they recommended three methods: classification, regression, and latent knowledge estimation. In the context of structured discovery, they prefer to use clustering, factor analysis, domain structure discovery, and network analysis [2]. Correlation mining, associate rule mining, sequential pattern mining, and fundamental data mining are most preferable methods. A few papers in LAK11[7] (Learning Analytics and Knowledge conference) highlighted several analytics approaches with innovative techniques: recommender systems improvement [46] cultural considerations in analytics [45], reputation mechanisms, and participating learning [12].

In higher education, Big data uses database systems that supply huge amounts of learners' data collected from variety learning and teaching activities. During the learning process learners leave data trails (streams) that make known their opinions, feelings, social influences, purposes, and objectives. Scientists can use these data to observe individual performance over learning process time.

[3] https://www.coursera.org/

[4] https://www.edx.org/

[5] https://www.udacity.com/

[6] https://www.thecompleteuniversityguide.co.uk

[7] http://lak15.solaresearch.org/

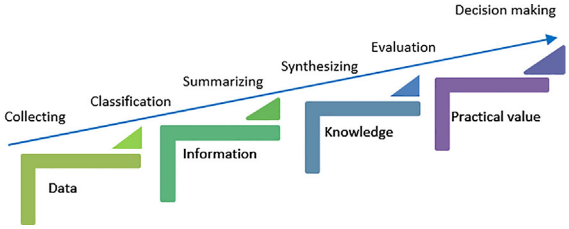

FIGURE 1 General process of mining insights from big data

In the rest of the paper, we will study the influence of big data in improving current challenges of education environments (see Figure 1). First we will classify the important worldwide tendencies in higher education institutions of and researches and present the potential of big data in addressing these changing progression. After that we will summarize opportunities associated to the implementation of big data and learning analytics in higher education. Various limitations and challenges of the current generation of big data techniques are also considered.

3 | BIG DATA TOOLS AND TECHNIQUES IN EDUCATION

Capability of conventional database systems is surpassing about the processing of big data, defined as primarily too big, and changes too fast [28]. Consequently, innovative tools and techniques are required to achieve, collect, store, manage, distribute, and explore larger sized data sets. The exploration of large amounts of data is not novel, however, big data rather states to evolving technologies which are intended for processing quantity dimensions' data of different types [10]. The tools and techniques are used by teachers, researchers, data managers, designers, administrators at the university.

Analysis in the big data environment essentially differs from analysis for small data. The general process of mining insights from big data [24] is shown in Figure 1. The deficiency of clearness about what precisely have to be measured in order to understand of how learning is being processed presents one of the large problems around learning analytics. Typical measurements cover number of logins, time spent, number of accessed resources, number of mouse clicks, number of finished coursework, etc.

The e-learning system is designed to allow access to the system at remote locations, providing anytime and anywhere access to course content, which is critical for promoting the use of e-learning systems [47] Therefore, we can identify three major capabilities of an e-learning system.

Functionality of an e-learning systems can be achieved by integrating numerous types of instructional and assessment media that are within the control of the system software or the learner [39]. Such audio, video, or text media allow learners to join course materials and solve homework tasks, solve tests, quizzes, and exercises online.

Interactivity as important features for the learning process define the communications among learners themselves or the collaborations among teachers and learners, and the interactions between learners and faculty [33]. Interactivity can be achieved uses common tools: e-mail, chat room, and bulletin board.

Response time. If the system has lowly response time it cannot be observed as functional or easy to use. Response time can be defined as the amount of time need for response which a learner perceives from the e-learning system [1]. It should be reliable, practical, and fast.

This section focuses on possibilities to properly use big data sets and data analytics in e-learning environments. We will present how data must be prepared for analysis and what we can learn about and from available data when apply analytic methods.

3.1 | Why associate big data and learning analytics together?

The process of gathering, bring together, and analyzing huge sets of data ("big data") useful for discovering useful information and some form of patterns is defined as big data analytics. Furthermore, big data analytics provides opportunities to better recognize the information which could be important to future decisions. The extraction of knowledge that comes from analyzing the data is of essential importance for big data analysts. According to Ferguson [18], the formal definition of learning analytics can be "the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs."

Big data concepts and analytics could be useful for a range of instructional and administrative applications in higher education: monitoring and checking learner performances, improving teacher's capacities, charges processing, commercial planning, donor finding [44]. The applications analyzed in this section will concentrate on learning and teaching, and specially observe process for extracting insights from big data. It is practically a condition that process of transaction be electronic in order to take benefit of big data and learning analytics. Traditional face-to-face processes can be useful for making decision, on the other hand, for the extensive and time-sensitive learning analytics applications, it is significant that instructional transactions are collected as they happen. It is possible for a course management/learning management system (CMS/LMS) and intelligent tutoring system. Most of such systems afford continual observing of learner activities, accesses to reading material, reactions, posts on a discussion board, solving of a test or quizzes, or some other evaluation. During the 15 week of online courses, it could be expected several thousands of transactions per learner. Recording and investigation of these transactions in real-time can be used as input data for learning analytics application. The general process of extracting practical values from big data can be broken down into five stages [24]:

1. Acquisition and recording;
2. Extraction, cleaning, and annotation;
3. Integration, aggregation, and representation;
4. Modeling and analysis;
5. Interpretation.

These five phases can be divided into two main sub-processes: data management and data/learning analytics. Based on our long lasting experiences in the field of e-learning and analysis of research studies implemented in educational and supplementary areas, we identify, recognize, and propose the widespread platform of making high-quality benefits that can be achieved by using big data techniques in the field of education (Figure 2). Figure 2 provides a detailed indication of all phases and presents a comprehensive overview of all the important factors that are required for conceptual and practical understanding of big data analytics within higher education. We propose this general model and hope that it will serve as inspiring starting point and bases for researchers in order to implement their own specific models and systems.

Educational performances and improvement an institution itself, regarding to resource distribution, learner

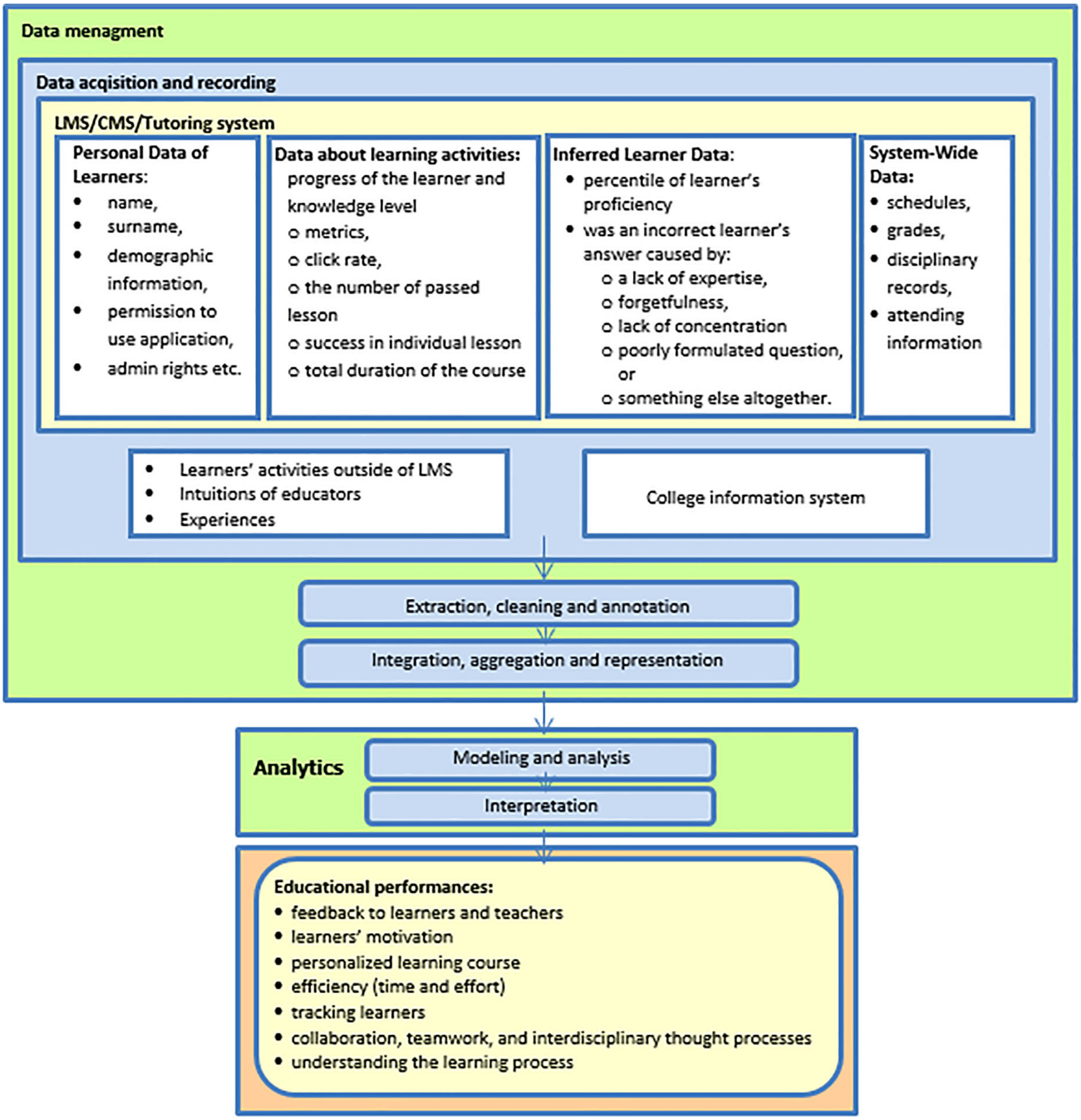
Data menagment
Data acqisition and recording
LMS/CMS/Tutoring system
Personal Data of Learners:
• name,
• surname,
• demographic information,
• permission to use application,
• admin rights etc.
Data about learning activities:
progress of the learner and knowledge level
o metrics,
o click rate,
o the number of passed lesson
o success in individual lesson
o total duration of the course
Inferred Learner Data:
• percentile of learner's proficiency
• was an incorrect learner's answer caused by:
o a lack of expertise,
o forgetfulness,
o lack of concentration
o poorly formulated question, or
o something else altogether.
System-Wide Data:
• schedules,
• grades,
• disciplinary records,
• attending information
• Learners' activities outside of LMS
• Intuitions of educators
• Experiences
College information system
Extraction, cleaning and annotation
Integration, aggregation and representation
Analytics
Modeling and analysis
Interpretation
Educational performances:
• feedback to learners and teachers
• learners' motivation
• personalized learning course
• efficiency (time and effort)
• tracking learners
• collaboration, teamwork, and interdisciplinary thought processes
• understanding the learning process

FIGURE 2 Big data analytical platform in higher education

achievement, administration, and finance are recognized. Classification of data types in the process of acquisition/recording plays an important role. Therefore, in this section, we will first explain what kind of data about learners can be identified according to the Knewton[8] study. Educational data can be divided into five types: four data sets about learner activities, one relating to learner identity, and invading. These types of data are obtained within the LMS/CMS/tutoring systems:

1. *Personality data*: These data include basic identification and descriptive information about the learners, such as: name, surname, demographic information, permission to use application, admin rights, etc. (descriptions of these data are in accordance with the answers to the following questions: Who are you? Do you have permission to use this application? What administration rights do you have? What region are you in? How about demographic info?)
2. *System-wide data*: These data include: schedules, grades, disciplinary records, and attending information. Teacher or principal can easy obtain within a class or school. It is not very useful at small scale according to little information about per-learner basis. It becomes more valuable at very large scale, and it may help inform system-wide recommendations.
3. *Conditional content data*: These data explain whether a part of content could be achieved through a group of learners. Also, these data are useful to identify what quantifiable learner knowledge improvements result, when a certain type of learner relates with a certain part of content. How well does a question actually assess what it intends to? Data on instructional materials involve algorithmically normed calculations. Therefore, it is not easy to generate these data.
4. *User interaction data*: These data include metrics about engagement, page views, click rate, bounce rate, etc. These metrics can advance user experience and maintenance, as the basis of internet optimization for customer web companies. This type of data is fairly easy to accumulate.
5. *Inferred learner data*: These data indicate the level of learners' knowledge about the concepts and percentile of their proficiency. Also, these type of data specify whether an incorrect learner's answer caused by a lack of expertise, a poorly formulated question, poor memory, lack of concentration, or something else. The answers to the question what is the possibility that the learner will permit the quiz next week, and how to improve the willingness for it, can be included in this type of data. Building of these data requires various content and a large number of learners, course designers, developers, and data researchers. Functional database architecture, tagging structure, innovative machine learning algorithms, and multifaceted taxonomic systems are required, as well.

It is very challenging for most educational institutions to build functionalities which achieve all five of the above mentioned data sets. Yet every institution must offer an answer for all five. The answer should be in the implementation of a general platform by using adequate solutions for each large data set. They should also be incorporated with data from the faculty information systems, learners' activities outside of LMS, as well as instructors' intuition, experiences, and instincts that can be used to improve strategies and procedures for succeeding courses of action.

3.2 | Benefits of big data and learning analytics

Big data offers new possibilities and new tasks for institutions of higher education. Siemens and Long [42] stated that big data presents the most studied framework in handling the huge collection of data and ping the future of higher education.

Big data signifies the exploration of a wide array of organizational and effective data, collected procedures designed at estimating formal performance and improvement of future performance prediction and recognizing prospective concerns related to learning, teaching, and research [21]. As well, some researchers specified that higher education has to include the analytics tool into the system in order to improve productivity. Several useful features of big data are estimated to help education in the near future [5,22,36].

1. *Feedback*: Response information and context perception can be useful for a big learning data. Learner often might fail at a subject but not know why (s)he is failing. It becomes valuable when the learner can look not just at himself, but at other people who have had the same experience. He/she can get an insight either that would describe it so (s)he is not frustrated or that (s)he could use to correct it so that (s)he could be successful again. The improvement of electronic learning modules supports evaluation of learners in logical, real-time ways. Data mining and data analytics software can provide immediate feedback to learners and teachers about educational performances. In order to predict learner outcomes such as dropping out, needing extra help, or being capable of more demanding assignments, this approach can analyze underlying patterns. Pedagogic approaches that seem most effective with particular learners could be identified.
2. *Motivation*: If the big data is appropriately implemented, learners possibly become committed in entering data to the process because they understand the power of how it works.

[8] http://www.knewton.com

3. *Tracking*: In order to understand the real patterns of learners more effectively Big data can be used for teachers, by allowing them to track a learner's experience in an e-learning course. In observing the digital paths learners leave overdue. Teachers are able to track learners' passage during the course of the whole learning experience.
4. *Collaboration*: Experts from many departments have to come together to retain a Learning Management System function at its best. This encourages cooperation, teamwork, and interdisciplinary thought processes.
5. *Efficiency*: Big data can save many hours of time and effort, when it comes to the achievement of our goals and strategies that we need to achieve them.
6. *Personalization*: Big data can be successful in the way we approach e-learning design by allowing designers to personalize courses to adjust their learners' individual needs. This will allow e-learning developers to promote the standard for effective and exceptional e-learning courses.
7. *Understanding the learning process*: By using big data in e-learning, teachers can see which parts of a course or exam were too easy and which parts were so difficult that the learner has failed to solve. Other parts of the learner's path teachers can analyze after that and consider pages reentered often, preferred learning styles, sections recommended to peers, and the time of day, when learning operates at its best.

   Despite the aforementioned educational performances, many faculties, and universities have confirmed that analytics can support significantly improvement an institution itself, regarding to resource distribution, learner achievement, administration, and finance. Some important features are listed below [15,42].

1. Improving administrative decision-making and organizational resource provision.
2. Innovating and transforming the college and university system, in addition to educational models and pedagogical approaches.
3. Assisting in creating common sense of complex topics through the combination of social networks and technical and information networks. Algorithms can recognize and provide insight into data and at risk challenges.
4. Helping leaders transition to holistic decision-making through analyses of "what-if" scenarios.
5. Exploring how different components within a complex discipline (e.g., remembering learners, decreasing costs) connects and discovers the influence of varying essential components.
6. Increasing administrative efficiency and productivity by providing latest information and allowing fast reaction to challenges.
7. Helping official leaders to control the hard (e.g., research, patent) and easy (e.g., quality of teaching, reputation, profile,) value created by faculty activities.
8. Evaluating typical grading techniques and instruments (i.e., departmental and licensing exams).
9. Testing and evaluation of curricula.

These features do not require learning analytics; nevertheless, increasing the amount, and kind of the collected data can be enhanced this approach significantly.

4 | THE APPLICATION DEVELOPMENT FRAMEWORK

Processing and analyzing big data requires special tools and systems that make available a computing environment to satisfy the requirements of analytics for large datasets. Therefore, we can conclude that for any big data platform there must be an application development framework which can make simpler the process of development, execution, testing, and debugging appropriate educational software components and building blocks. Such type of framework should contain [27]:

1. model development tools and techniques;
2. capability for program loading, implementation, and process scheduling;
3. the system configuration and management tools.

There are several technologies to handle big data such as Map-Reduce [17], Hadoop,[9] NoSQL [20], PIG [31], and Hive [43]. Relational databases propose a flexible manner to collect, save, and operate data using a Structured Query Language (SQL) which is easy to study and very suitable for managing medium-sized datasets [9]. However, standard relational databases come to be impractical for several gigabytes of data or several million explanations. Databases for managing large amount of data are generally known as NoSQL databases [20]. The term can be understood as connotation "not only SQL." NoSQL databases are simpler than SQL databases regarding data handling abilities. A lot of scientists have established it is obligatory to develop systems which can analyze and process billions of transactions per day. The need for processing data sets of this size has led to the enlargement of several tools for analysis and management of Big Data.

[9] http://www.cloudera.com/content/cloudera/en/about/hadoop-and-big-data.html

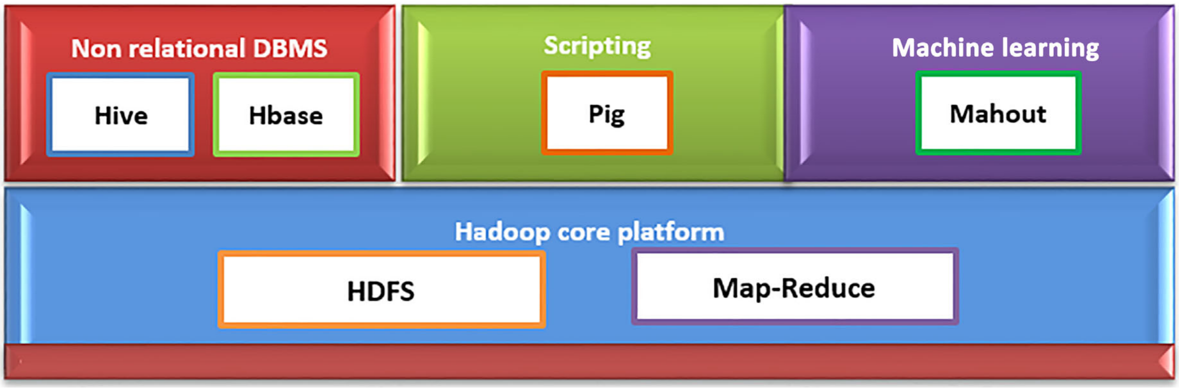

FIGURE 3 Hadoop conceptual framework

One of the most known and significant tools commonly used by most data management systems is Hadoop. Hadoop is an open-source software framework, written in Java, intended for distributed storage, and processing of very large data sets, using simple programing models. Principally, it achieves two tasks: large data storage and faster processing. In the Figure 3, we can see Hadoop conceptual framework. Hadoop is responsible for the strong Hadoop Distributed File System (HDFS), encouraged by Google's file system, as well as parallel programing model using the Map-Reduce paradigm [17]. Therefore, program execution is divided into a Map and a Reduce phases, separated by data transfer between nodes in the cluster [3]:

1. Map phase. A node performs a Map function on a segment of the input data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). The records for any given key, possibly spread across many nodes, are grouped at the node running the Reducer for that key. This involves data transfer between machines.
2. Reduce phase. This second Reduce stage takes the output from a Map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name Map-Reduce implies, the reduce task is always performed after the map job.

This is a simple programming model, constrained to use of key-value pairs, but large numbers of tasks and algorithms will adjust into this framework. In addition to Hadoop itself, there are multiple open source projects built on top of Hadoop. Major projects, adapted for application in the education field, are described below.

Hive [43] is a data warehouse framework constructed on top of Hadoop. Hive supports analysis of large datasets stored in Hadoop's HDFS. Hive can be used for ad hoc querying with an SQL type query language in order to create reports, summaries, analyzes, evaluation results, etc. Hive includes different storage types. It desires introduction of Map-Reduce jobs. It is aimed for batch processing, not online transaction processing. It does not propose real-time queries.

Pig [31] is a platform for evaluating large data sets that consists of a high-level language (Pig Latin) and implementation framework whose compiler produces series of Map-Reduce programs for completing within Hadoop. Pig's infrastructure layer has the following features:

1. Ease of programming. Complex tasks consist of numerous data transformations which are obviously encoded as data flow sequences, making them easy to write, understand, and maintain.
2. Optimization opportunities. The way in which tasks are encoded permits the system to optimize their execution automatically, allowing the user to focus on semantics rather than efficiency.
3. Extensibility. Pig users can construct particular functions to meet their specific processing requirements.

Pig is a Java client-side application, and leaners install locally—nothing is altered on the Hadoop cluster itself. Grunt is the Pig interactive shell. Pig is designed for batch processing of data. These Hadoop components are usually grouped together but you can replace each component, or add new ones, as desired.

Mahout [32] is a library of scalable machine-learning algorithms constructed on top of Hadoop and using the Map-Reduce paradigm. Machine learning can be defined as an artificial intelligence discipline which is aimed at supporting the machines to learn without being explicitly programed. Current algorithm emphasis following areas of Mahout [32]:

1. *Collaborative filtering*—tracks learner behavior, searches for similar learners, and makes learning resources recommendations.
2. *Clustering*—takes items in a particular class (web pages, learning objects, or lessons) and categorizes them into

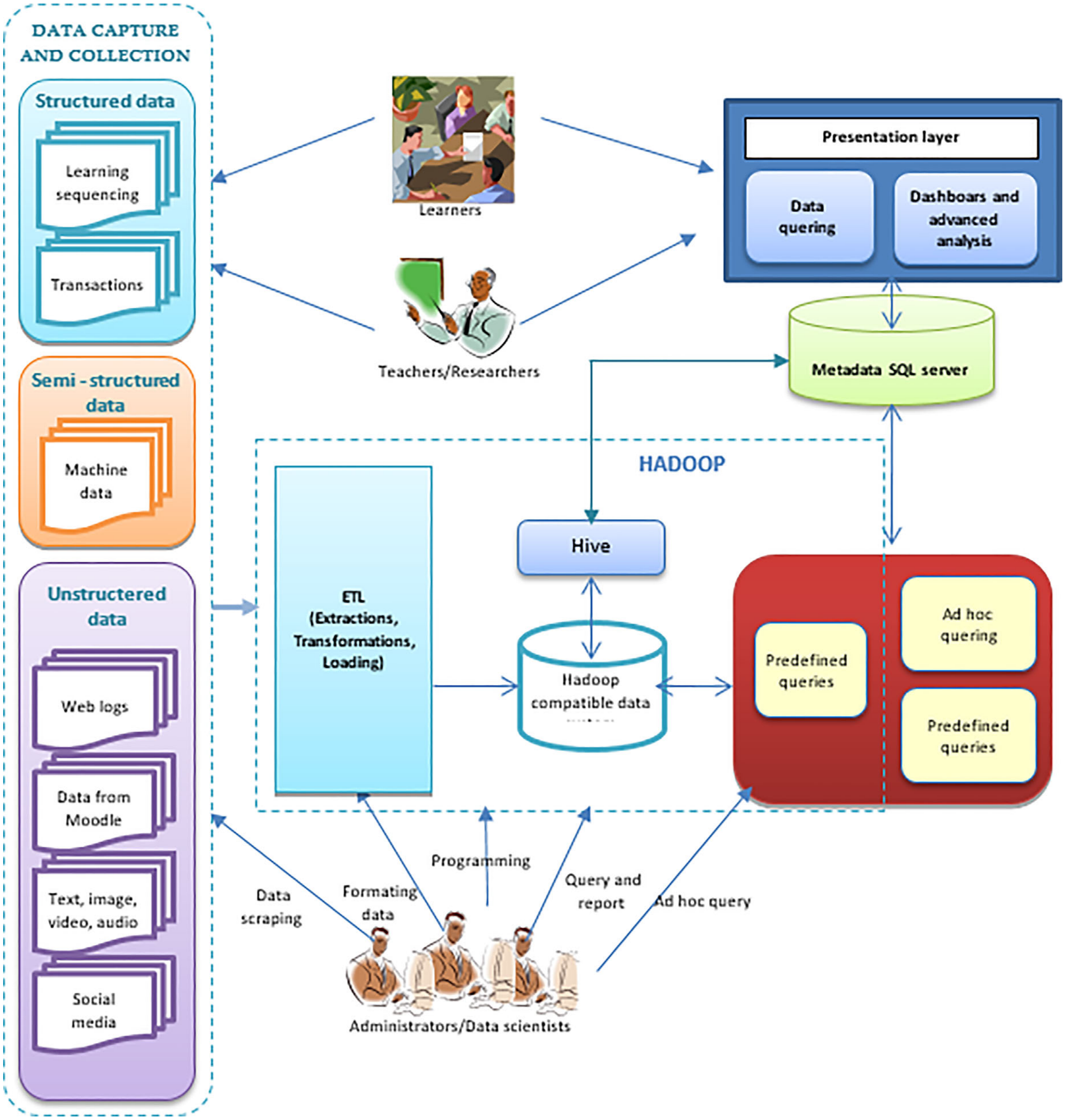

FIGURE 4 Architecture framework for support higher educational research

logically organized groups, such that items fit in the same group are similar to each other.

3. *Classification*—learns from existing categorizations and then appoints unclassified items to the best category.
4. *Frequent item set mining*—analyzes items in a group (e.g., lessons or objects in learning session) and then recognizes which items typically appear together.

HBase[10] is a distributed, scalable, big data store that runs on top of HDFS. HBase enables real-time and random access to large data. HBase can be used for hosting very large tables. Tables in HBase can serve as the input and output for Map-Reduce jobs run on Hadoop. HBase is very different from traditional relational databases, like MySQL, PostgreSQL, or Oracle in how it's architected and the features that it provides to the applications using it. HBase provides many features that can be grouped as a column family, so that the elements of the family of the column all stored together. This differs from the row oriented relational database, where all the columns of a certain row are kept together. HBase needs the redefined table and specific families of columns. On the other hand, the schema is flexible and adjustable to changing application requirements. New columns can be added to families at any time.

[10] http://www.thecompleteuniversityguide.co.uk/distance-learning/moocs-%28massive-open-online-courses%29/

The architectural design of ecosystem that can support learning process in-higher education is very challenging and complex. According to our analysis in previous section, it can be observed that higher educational data contains complex associations, a very large number of variables and parameters to be considered, and has the potential to become big data. The general principles of system design undertake the following requirements [48]:

1. Support for large volumes and multi-structured data sets.
2. Platform independent deployment on the learner side.
3. An easy-to-use learner interface.
4. Incorporated analytic modules that allow learners to complete course quickly and answer their own questions.
5. Embedded statistical functions as well as allow custom statistical codes.
6. User-friendliness, high-level flexibility, and scalability by using high performance computing and cloud computing resources.

On the basis of the conclusions, we reached in the study of big data and learning analytics in the education field, we propose one principal and general architectural framework to support higher educational research in these areas. This is a good example of ecosystem that supports multi-structured data sets processing and analyzing (Figure 4). The roles and abilities of learners, researchers, teachers, and data scientists are clearly defined and proposed within this framework. The framework, we are proposing in Figure 4, is composed of five modules:

1. *Data capture and collection module* maps, aggregates and cleans data from different sources and prepares data for ETL (Extractions, Transformations, and Loading) process. Structured data, which constitutes only 5% of all existing data [14], refer to the tabular data found in spreadsheets or relational databases. In the case of education, such type of data can be in the form of learning sequencing, transactions, etc. Text, images, audio, video, and data from Moodle are examples of unstructured data, which sometimes lack the structural organization required by machines for analysis. Spanning a continuum between fully structured and unstructured data, the format of semi-structured data does not conform to strict standards. Extensible Markup Language (XML), a textual language for exchanging data on the Web, is a typical example of semi-structured data [19]. XML documents contain user-defined data tags which make them machine-readable.
2. *The ETL module* includes three main functions: data integration to relevant tables, data transformation, and loading of data specifically for analysis. The ETL process populates the databases with data that is analysis ready, allowing more confident analysis.
3. The different data sources are managed in the *Hadoop platform* and are stored in a compatible repository. The Hadoop platform allows easy management of high volume and various data.
4. *The analysis engines* execute standard and predefined procedures in order to enable complex statistical analysis.
5. *The presentation layer* provides a user-friendly graphical interface. Learners, teachers, and researchers could easily retrieve information via this interface without the need for the in-depth data analysis knowledge, programming skills, or database schema background.

We made this proposal, according to the most important aspects of the implementation of big data and analytics in the field of education: integration data from different sources, embedded statistical environment, high performance computational back-end, and the possibility of high-quality presentation layer to the end user: learners, teachers, and researchers. The presented architecture solution can be useful for providing exceptional high performance and new, flexible data processing framework for ongoing study in higher educational research.

5 | THE CHALLENGES OF IMPLEMENTATION BIG DATA TECHNIQUES IN HIGHER EDUCATION

Regardless of the fast development of applications that support implementation of big data techniques in higher education, there are also a number of apprehensions that need to be considered. Becker [5] suggested three cooperative elements to be examined during the collection of data for analytics: timing, population, and location. Any time unit, starting from the second until semester or year can be defined as timing element. The features of the group of students participating in the learning environments refer to the population. According to the learning space where students are retrieving the knowledge, the location could be specified.

Some researchers have been recognized that are important for learners to be able to critically analyze resources and information. effectively direct their own learning in an open online networked environment. Innovative facilities and an inspired mindset in an environment that is characterized by change and complexity have also been highlighted as important [16,38].

One of the major barriers for implementing big data analytics in higher education could be financial expenses [8]. Many institutions view analytics as an expensive effort rather than as an investment. Much of the concern around affordability focuses on the supposed requirements for expensive methods of data collection.

There is a need for more investment in analytics professionals, competent to use big data, and analytics properly. They should be able to monitor the entire process, from defining the important questions to developing data models for designing and delivering alerts, recommendations, and reports. Furthermore, knowledgeable designers and proficient database administrators, accomplished with warehousing, and incorporating data through numerous files and formats, are required [8]. Besides the expertise required for database development, istructional designers who work with the university will have to understand learner behaviors that are appropriate for the application at hand. Good knowledge of statistics, decision trees, and strategy mapping are also important for the development of an algorithm to build prediction models.

The main issues that can be observed in the application of the use of big data in education are relate to data profiling, privacy, and the rights of learners with respect to their individual behavior recording [7]. Despite the fact that traditional classes approach always evaluates performances and academic behavior of learners, learning analytic is useful in a process of tracking students' behavior on a completely different level and scale, and should be assessed. Despite the fact that learning analytics may be useful concerning learners' success, the "big data" approach might as well be seen as an offensive of privacy that some learners would rather not have enforced upon them. For example, some important issues must be taken into consideration [36]: should learners be expressed that their activity is being followed? How much information is necessary for faculty, students, parents, scholarships, and other issuers? In what manner should faculty affiliates act in response? Do learners have a requirement to look for support? Protection should be undertaken to confirm that the well-known collections of personal data of learner transactions are not hurt in a way that probably hurt individuals.

One encouraging approach to resolving these issues could be masking the data at its source [4]. Masking is one type of creative approaches that will make large-scale applications of data possible while still protecting the confidence of students' and teachers' information. New performances and competences for ETL software applications permit sensitive data to be masked at the database level, when brought into a data warehouse. It can be concluded that even if someone has acquired physical access to database, delicate information like social security numbers will still be confused.

6 | CONCLUSIONS

Big data have a significant impact on higher education, practice, from improving learners experience, and knowledge through enhanced academic studying, to more operative decision making, and to planned response to varying overall trends. Big data have the potential to address some of the crucial challenges facing the current higher education. Big data give good grounds for expecting to turn complex, often unstructured data into actionable information. Moreover, big data provide scientists with the opportunity to comprehend the meanings of these data and how they can be analyzed in a significant, effective, and consistent manner that contributes to both theory and practice. For that reason, all of education organizations need to have a strategy for how it will take advantage of big data.

In this paper, we described big data concept, possibilities, techniques, and tools which can be used for extending the capabilities of educational systems. We highlighted the significance of big data and analytics in education, which is twofold: in managing reform activities in higher education, and assisting instructors in improving teaching and learning. Various challenges of the current generation of big data systems are considered. This research as well observes theoretical and evolving framework which can support data organization and integration of educational research in addition to develop key performance displays and techniques for collecting and processing large amounts of data. In addition, a set of the potential privacy, ethical, and data profiling questions that might occur from using academic data to predict success of students in higher education are being tracked.

In the future, educational institutions should try to balance the institutions own philosophy of learner development on the one side and various federal privacy laws on the other side. It is significant that organizations comprehend the dynamic nature of educational success and retaining, offer surroundings for open dialogue, and enhance practices and approaches to solve these issues. Interesting additional features, which are worth for further research, can be: optimizing the presented architecture with other learning services, identifying additional tactics, automated them, and integration with cloud platform.

We hope that research achievements presented in this paper and the implications derived from them will advance the discussion about building effective learning systems, based on big data, and learning analytics, for learners, instructors, course designers, and institutions.

REFERENCES

1. J. Bailey and S. W. Pearson, *Development of a tool for measuring and analyzing computer satisfaction*, Manage. Sci. **29** (1983), 530–545.
2. R. Baker and J. D. Siemens, Learning, schooling, and data analytics. *Handbook on innovations in learning for states, districts, and schools*. Center on Innovations in Learning, Philadelphia, PA, 2013, pp. 179–190.

3. K Bakshi, *Considerations for big data: Architecture and approaches*, In Proceedings of the IEEE Aerospace Conference. 2012, pp. 1–7.
4. M. Barlow, *Real-time big data analytics: Emerging architecture, O'REILLY*, Kindle edition, 2013.
5. B. Becker, *Learning analytics: insights into the natural learning behavior of our students*, Behav. Soc. Sci. Librarian, **32** (2013), 63–67.
6. M. Bienkowski, M. Feng, and B. Means. Enhancing teaching and learning through educational data mining and learning analytics, (2012), 45.
7. D. Boyd, Privacy and publicity in the context of big data. Raleigh, North Carolina, (2010).
8. J. P. Campbell, P. B. DeBlois, and D. G. Oblinger, *Academic analytics: A new tool for a new era*, EDUCAUSE Review, **42** (2007), 40.
9. R Cattell, *Scalable SQL and NoSQL data stores*, ACM SIGMOD Record, **39** (2011), 12–27.
10. H. Chen, R. H. Chiang, and V. C Storey, *Business intelligence and analytics: From big data to big impact*, MIS quarterly, **36** (2012), 1165–1188.
11. M. Chen, S. Mao, and Y. Liu, Big data: A survey, mobile networks and applications, **19** (2014) 171–209.
12. D. Clow and E. Makriyanis, *iSpot analysed: Participatory learning and reputation*, In Proceedings of the 1st International Conference on Learning Analytics and Knowledge, (2011) 34–43.
13. S. Cooper and M. Sahami, *Reflections on stanford's moocs*, Commun. ACM, **56** (2013), 28–30.
14. K. Cukier, The Economist, data, data everywhere: A special report on managing information, February 25, 2010 Available online at https://www.economist.com/node/15557443
15. B Daniel, *Big data and analytics in higher education: Opportunities and challenges*, Brit. J. Educ. Technol, **46** (2015), 904–920.
16. S. Downes, *New tools for personal learning*, Procedings of the MEFANET Conference, Czech Republic, (2009).
17. EMC: Data science and big data analytics. *EMC education services*. 2012, pp. 1–508.
18. R. Ferguson, *Learning analytics: Drivers, developments and challenges*, Int. J. Technol. Enhan. Learn. **4** (2012), 304–317.
19. A. Gandomi and M. Haider, *Beyond the hype: Big data concepts, methods, and analytics*, Int. J. Inf. Manage. **35** (2015), 137–144.
20. R. Hecht and S. Jablonski, Nosql evaluation. In *International conference on cloud and service computing*. 2011, pp. 336–341.
21. F. A. Hrabowski, III and J. Suess, *Reclaiming the lead: Higher education's future and implications for technology*, EDUCAUSE Review, **45** (2010), 6. Available online at https://www.educause.edu/library/ERM1068
22. F. A. Hrabowski, III, J. Suess, and J. Fritz, *Assessment and analytics in institutional transformation*, Assess. Analyt. Institut. Transform. **46** (2011), 14–28.
23. L. Johnson et al. NMC horizon report: 2014 K, (2014), 1–52.
24. A. Labrinidis and H. V. Jagadish, *Challenges and opportunities with big data*, Proc. VLDB Endowment, **5** (2012), 2032–2033.
25. D. Laney, *3D data management: Controlling data volume, velocity and variety*, Gartner, (2015). Meta Group. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controllin g-Data-Volume-Velocity-and-Variety.pdf. Accessed 16 Jan 2016.
26. L Leydesdorff and H. Etzkowitz, *The transformation of university–industry*–government relations, Electron. J. Soc. **5** (2001), 1–17.
27. S. Madden, *From databases to big data*, IEEE Internet Comput, **16** (2012), 4–6.
28. J. Manyika et al. Big data: The next frontier for innovation, competition, and productivity, (Accessed 3 January 2015), (2011).
29. J. Needham, *Disruptive Possibilities: How big data changes everything, O'REILLY* Kindle edition, 2013, p 90.
30. New Media Consortium. Horizon report: 2012 higher education edition. [2013-10-03], (2012), Available online at https://www.net.edueause.edu/ir/librarylpdf/HR2013.pdf
31. C. Olston et al. *Pig latin: a notso-foreign language for data processing*, In: SIGMOD '08: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, ACM, Vancouver, Canada, (2008), 1099–1110.
32. S. Owen et al. Mahout in action, Manning Publications Co. (2011).
33. M. Palloff and K. Pratt, *Building learning communities in cyberspace*. Jossey-Bass Publishers, San Francisco, 1999.
34. Z. Papamitsiou and A. A Economides, *Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence*, J. Edu. Technol. Soc. **17** (2014).
35. L. Pappano, *The year of the MOOC*, The New York Times (2012). Retrieved from: http://www.nytimes.com/2012/11/04/education/edlife/massive-open-onlinecourses-are-multiplying-at-a-rapid-pace.html?pagewanted=1
36. A. G. Picciano, *The evolution of big data and learning analytics in American higher education*, J. Asynchron. Learn. Net. **16** (2012), 9–20.
37. J. Reich, *Rebooting MOOC research*, Science, **347** (2015), 34–35.
38. P. Sahlberg, *Creativity and innovation through lifelong learning*, Lifelong Learn. Eur. J. **14** (2009), 53–60.
39. B. Seels and Z. Glasgow, *Making instructional design decisions*. Upper Saddle River, NJ:Prentice-Hall, 1998.
40. M. Sharples et al. Innovating Pedagogy, (2014), 1–37.
41. G. Siemens, *Learning analytics: envisioning a research discipline and a domain of practice*. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, (2012), 4–8.
42. G. Siemens and P. Long, *Penetrating the fog: Analytics in learning and education*, EDUCAUSE review, **46** (2011), 30.
43. A. Thusoo, et al. *Hive: A warehousing solution over a map-reduce framework*, Proceedings of the VLDB Endowment, **2** (2009), 1626–1629.
44. A. van Barneveld, K. E. Arnold, and J. P. Campbell, *Analytics in higher education: Establishing a common language*, EDUCAUSE Learn. Initiat. **1** (2012), 1–11.
45. R. Vatrapu, *Cultural considerations in learning analytics*. In Proceedings of the 1st International Conference on Learning Analytics and Knowledge, (2011), 127–133.
46. K. Verbert et al. *Dataset-driven research for improving recommender systems for learning*. In Proceedings of the 1st International Conference Learning Analytics & Knowledge, (2011), 44–53.
47. M. Waheed and K. Kaur, Review of e-learning knowledge quality dimensions: Concepts and measurements. (2014).

48. P. Xuan et al. An infrastructure to support data integration and curation for higher educational research. In *Proc. BT 8th IEEE international conference on e-Science*. Chicago, 2012, pp. 8–12.

A. Klašnja-Milićević is an assistant professor at Faculty of Sciences, University of Novi Sad, Serbia. She joined the graduate program in Computer Sciences at Faculty of Sciences, Department of Mathematics and Informatics, University of Novi Sad in 2003, where she received her MSc degree (2007) and PhD degree (2013). Her research interests include e-learning and personalization, Intelligent Tutoring Systems, information retrieval, Internet technologies and recommender systems. She actively participates in several international projects. She has also served as Program Committee member of several international conferences. She co-authored one university textbook and one monograph. She has published over 30 scientific papers in proceedings of international conferences and journals.

M. Ivanović holds the position of full professor at Faculty of Sciences, University of Novi Sad. She is a member of the University Council for Informatics. She is author or co-author of 13 textbooks, several monographs and more than 330 research papers on multi-agent systems, e-learning, and intelligent techniques, most of which are published in international journals and conferences. She is/was a member of Program Committees of more than 200 international conferences, leader of numerous international research projects. She was principal investigator of more than 15 projects. Mirjana Ivanovic delivered several keynote speeches at international conferences, and visited numerous academic institutions all over the world as visiting researcher and teacher. Currently she is Editor-in-Chief of the Computer Science and Information Systems journal.

Z. Budimac holds the position of full professor at Faculty of Sciences, University of Novi Sad, Serbia. Currently, he is head of Computing Laboratory and Chair of Computer Science. His fields of research interests involve: Software Engineering, Programming Languages and Tools, Educational Technologies, Agents and WFMS, Case-Based Reasoning. He was principal investigator of more than 20 projects. He is author of 13 textbooks and more than 250 research papers most of which are published in international journals and international conferences. He is/was a member of Program Committees of more than 100 international Conferences and is member of Editorial Board of Computer Science and Information Systems Journal.

How to cite this article: Klašnja-Milićević A, Ivanović M, Budimac Z. *Data science in education: Big data and learning analytics*. *Comput Appl Eng Educ*. 2017;25:1066–1078. https://doi.org/10.1002/cae.21844