Chapter

Improving Online Education Using Big Data Technologies

Karim Dahdouh, Ahmed Dakkak, Lahcen Oughdir and Abdelali Ibriz

Abstract

In a world in full digital transformation, where new information and communication technologies are constantly evolving, the current challenge of Computing Environments for Human Learning (CEHL) is to search the right way to integrate and harness the power of these technologies. In fact, these environments face many challenges, especially the increased demand for learning, the huge growth in the number of learners, the heterogeneity of available resources as well as the problems related to the complexity of intensive processing and real-time analysis of data produced by e-learning systems, which goes beyond the limits of traditional infrastructures and relational database management systems. This chapter presents a number of solutions dedicated to CEHL around the two big paradigms, namely cloud computing and Big Data. The first part of this work is dedicated to the presentation of an approach to integrate both emerging technologies of the big data ecosystem and on-demand services of the cloud in the e-learning field. It aims to enrich and enhance the quality of e-learning platforms relying on the services provided by the cloud accessible via the internet. It introduces distributed storage and parallel computing of Big Data in order to provide robust solutions to the requirements of intensive processing, predictive analysis, and massive storage of learning data. To do this, a methodology is presented and applied which describes the integration process. In addition, this chapter also addresses the deployment of a distributed e-learning architecture combining several recent tools of the Big Data and based on a strategy of data decentralization and the parallelization of the treatments on a cluster of nodes. Finally, this article aims to develop a Big Data solution for online learning platforms based on LMS Moodle. A course recommendation system has been designed and implemented relying on machine learning techniques, to help the learner select the most relevant learning resources according to their interests through the analysis of learning traces. The realization of this system is done using the learning data collected from the ESTenLigne platform and Spark Framework deployed on Hadoop infrastructure.

Keywords: online education, e-learning, online learning, big data, cloud computing

1. Introduction

Firstly, distance education has been developed mainly to make distance education accessible to all those who cannot follow face-to-face education, whether for reasons of geographical distance, lack of financial resources, or for lack of time.

In fact, since the emergence of programming languages, several learning systems have been created that aim to computerize the pedagogical activity in order to facilitate learners' access to training without having to be physically present. So, e-learning platforms are mainly computer applications dedicated to education [1]. The Francophone community, which is interested in the development of teaching through computer technologies, calls all systems, platforms or software for learning by Computing Environments for Human Learning (CEHL). In this context, a CEHL is a set of tools, systems and learning platforms based on the use of new information and communication technologies to enable to the learner to realize his formation course without constraints of time and distance. They are designed to promote learning and build knowledge for the learner through computer-based learning situations in the form of classes, exercises, or other activities. Many terms have been used to describe this mode of distance learning, including E-learning, online learning, CAT (Computer Assisted Teaching), ITS (Intelligent Tutorial Systems), Distance Education, Distance Training, WBT (Web-Based Training), M-learning (Mobile Learning), MOOC (Massive Open Online Courses), etc. [2]. As e-learning has many advantages such as flexibility, diversity, openness, etc., it becomes an essential way to acquire new knowledge and skills and to take lessons anytime, anywhere and from any device. And, this will not replace face-to-face teaching but greatly enhances the effectiveness of education.

Today, the IT world is experiencing a strong technological development in terms of resource acquisition, data management, and manipulation. This development is marked by the evolution of the cloud computing model and the advent of the new generation of big data technologies. All of these technologies have upset existing practices by introducing new forms of on-demand services, scalable architectures, and distributed approaches for distributed data processing and analysis.

In this context, the cloud aims to meet the strong demand for e-learning from universities and training organizations that need to develop, execute and deploy high-performance applications at an optimized cost. It also aims to enable e-learning professionals to deal with the multiplicity of devices (desktop, smartphones, and tablets), operating systems, programming languages, and development frameworks. Furthermore, big data technologies will allow an e-learning system to automatically evolve and adapt to different situations depending on the learner's profile and interactions. Indeed, thanks to the integration of big data in CEHL, an online learning system now has the ability to make decisions and make predictions automatically, without the intervention of a human being, through very advanced models and algorithms of machine learning, which is an integral part of big data.

Recent IT innovations have transformed all areas, including the field of distance learning. Indeed, CEHL have always evolved. Since, the computer, at first, the Internet, then, the cloud computing and the big data, at the moment, appeared for educational use. However, these CEHLs have never been perfect. They are constantly asking for improvements and adjustments simply because we are facing many changing realities and technological developments. So systems, structures, and e-learning processes must remain flexible to be able to adjust to these changes. Our work is part of the research work on CEHL and focuses on the integration of new information and communication technologies into distance education. More specifically, this chapter situates in the context of the development and continuous improvement of CEHL through the implementation of an open, adapted and intelligent online learning platform, which takes into account dimensions of resource sharing, availability and quality of the learning service, and at a lower cost.

To do this, it has proved essential to incorporate modern technological innovations into the CEHL. Thus, this work chooses to use the services provided by cloud computing, as a promising IT model, enabling the outsourcing of hardware,

software, and platform resources to remote servers of a cloud provider in order to make them accessible on the internet in the form of services while facilitating their acquisition and optimizing their exploitation. Moreover, it uses the latest technological advances through the implementation of advanced technologies of the big data ecosystem in e-learning systems. Big data offers, in addition to massively parallel computational powers and distributed storage capabilities, sophisticated methods, and algorithms dedicated to machine learning to process and extract knowledge from the various types of data produced by e-learning systems, including learner profile information, activities, preferences, results, etc. This work set up a recommendation engine, the world's most popular big data application, to help and guide the learner to easily identify and select the most relevant educational resources by offering him an intelligent system capable of generate a catalog of courses adapted to their interests and their cognitive level through the analysis of historical data of learning activities.

2. Related works

The integration of big data and the cloud into e-learning systems is one of the main tracks of this research work which aims to take advantage of on-demand services, and exploit powerful technologies of distributed storage, parallel processing, and real-time analysis to solve some problems of e-learning systems, e.g. recommendation, prediction, motivation, and handle huge amounts of heterogeneous data. This section presents the state of the art of big data and cloud in relation to the e-learning domain. It explores the research that has been done by researchers to integrate these three great paradigms by proposing architectures, approaches and use cases.

In this context, the work [3] addresses the incorporation of the big data into the online learning system by proposing a framework to provide a high quality learning service. This framework facilitates the combining of the e-learning field and big data ecosystem in order to benefit from the advanced techniques of data management and analysis. It consists of three layers. Firstly, there is the e-learning layer which includes educational pedagogical methods, teaching contents, and a set of technologies dedicated to education such as Learning Management System (LMS), Virtual Class Room (VCR). This layer also contains learner information, e.g. profile, preferences, interests, and interactions with the learning system. Secondly, there is the big data layer. Indeed, the data collected from the e-learning platform are passed through this extraction and processing layer in order to prepare, to analyze this data which will be transmitted, subsequently to the third layer of results interpretation and data visualization via several presentation techniques provided by dedicated user applications.

This work [4] aims to show how big data can help solve some distance learning problems by exploiting its technologies in learning content analysis to ensure efficiency and reliability e-learning systems. It uses some powerful big data techniques and tools, such as Hadoop, MapReduce, and HDFS. It also aims to propose a methodology to incorporate the tools and Frameworks big data in the field of CEHL. This methodology includes four steps, which are: first, identifying likely sources of educational data. These data can be of various types such as e-learning databases, e-mails, social networks, etc. Then the third step is data extraction which concerns the collection of data from different sources. Next, there is the big data processing step, which consists in choosing the most appropriate programming language and identifies an appropriate algorithm for returning the results requested by the user. The visualization of data is the fourth step that aims to present the results of analyzed data in an interactive visual form for accessible and understandable.

The research work [5] addresses the use of big data in the academic context. It proposes a model for adapting big data technologies with e-learning platforms. This aims to integrate existing LMSs already in universities with the Hadoop Framework deployed as a cloud SaaS service. In order to exploit the data of a traditional LMS system, Hadoop uses two methods for this purpose. The first method is to migrate the relational database from LMS to an HDFS distributed file system. The second mechanism is to transfer the structured data from LMS into a Data Warehouse. After integrating LMS data into the Hadoop Framework, it is possible to apply conversions and filtering to this data, and then perform advanced analysis. These analyses can greatly contribute to the development of an adaptive and personalized learning system.

We note the research work that was done in this field does not provide a real use case or application based on the software libraries of big data ecosystem in online learning systems, such as high performance machine learning techniques of Spark MLlib or scalable algorithms of Mahout Framework. This is why in this work we were interested, firstly, in proposing a new approach for integrating big data, cloud computing and online learning systems. In addition, this chapter attempts to set up a large-scale big data application. It consists in developing a recommendation system capable of providing adapted and personalized courses to each learner according to his preferences, his cognitive level and his learning style.

3. Big data

In the literature, the term big data first appeared in 1997 according to the IEEE digital library archives, in a published scientific article [6] by two NASA researchers: Michael Cox, and David Ellsworth, on the technological challenges of visualizing large data sets and the difficulty of systems in dealing with massive volumes of data.

In 2003, Google has published a paper about the idea behind of its file system [7], and reveals the first secrets of the success of its search engine. One year later, Google developed MapReduce as a parallel and distributed computing model programming for massive data processing. A year later, Doug Cutting and Michael Cafarella, at that time employed at Yahoo and inspired by the principle of MapReduce, develop Nutch Search Engine, which will become today Apache Hadoop [8].

3.1 Defining big data

According to International Data Corporation (IDC), "Big Data Technologies describes a new generation of technologies, architectures, tools, and techniques designed for extracting value from very large volumes of a wide variety of data, allowing a high speed of capture, discovery and/or analysis " [9]. Gartner in his report [10] gave the following definition: "big data brings together data of great variety, arriving in increasing volumes, at high speed. This is called 3V". In other words, big data is composed of complex and very diverse data. This large data is generated at a high speed as current database systems become unable to handle it. So, big data tools can provide effective solutions to overcome these challenges. NIST suggests that, "Big data is when data volume, acquisition speed, or data representation limits the ability to perform efficient analysis using conventional relational models or requires the use of a significant horizontal scale for effective handling. big data refers to the need to distribute and parallelize the data computing and storage in data-intensive applications" [11].

Specifically, big data can be divided into data science and big data technologies. Data science is "the study of techniques covering the acquisition, conditioning,

evaluation and exploitation of data", while big data technologies are "systems, software libraries, tools, Frameworks with their algorithms associates that allow distributed processing and analysis of big data problems between clusters of machines" [9].

3.2 Big data characteristics

The characteristics of big data are five, or the 5V model. Vs refer to five key elements that are: volume, variety, velocity, veracity, and value. **Figure 1** gives a summary of the different characteristics of big data.

Volume refers to the huge amount of data to be stored, processed, analyzed and disseminated by big data tools and technologies. Indeed, the volume of data, generated and handled by companies, is constantly increasing. Currently, the data is measured in petabytes, exabytes, even zettabyte.

Variety refers to the variety of formats and types of data. In fact, big data technologies can handle heterogeneous data from various sources. The classic format is that of the relational database, in which the data is stored according to a rigid and organized schema. But currently, more than 80% of the data generated by companies is of the semi-structured and unstructured type, for example, text, image, video, voice, etc. For this, big data offers the ability to gather all these data and analyze them.

Velocity refers to the speed or frequency at which the data are generated and used. This aggregated data must be exploited in real time. This requires high-performance computing and storage powers and robust analysis tools. In this sense, the IaaS services of the cloud prove to be an adequate solution allowing theoretically unlimited computing resources.

Veracity means the validity and quality of the data captured. It is the credibility and reliability of the data on which a data scientist is based to perform analysis in order to make decisions. Therefore, the big data platform and solutions aim to select and search the exact data in giant databases by eliminating useless data, through innovative tools and techniques [12].

Value refers to the ability of big data to derive value from huge masses of heterogeneous data. It's good to have access to large volumes of data, but we still have to

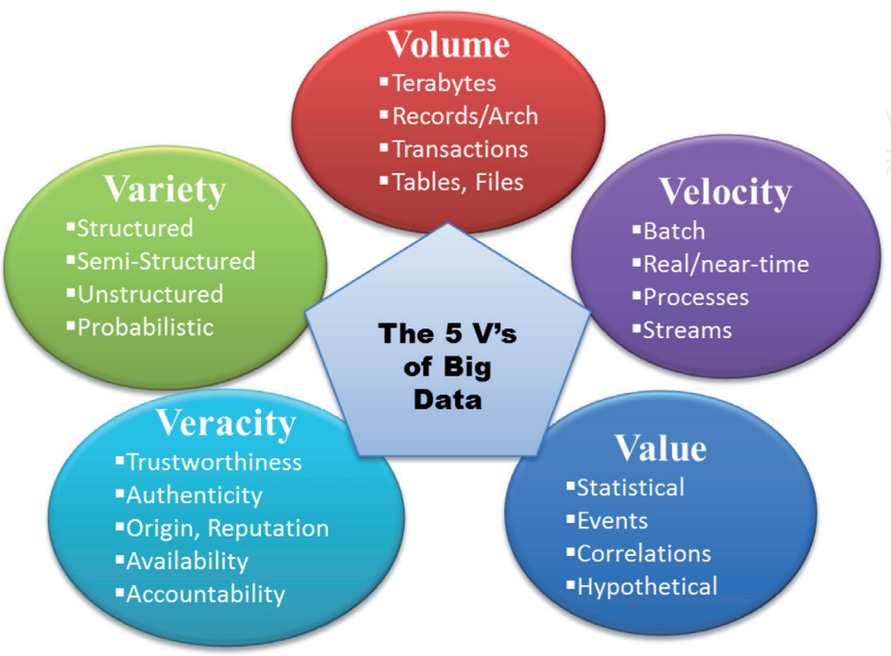

Figure 1.
Characteristics of big data or 5V.

turn them into value. Indeed, it is necessary that the new generation of distributed technologies (Hadoop, Spark, etc.), DBMS NoSQL, and advanced methods, serve something useful and usable for businesses and universities.

3.3 Types of data

In general, there are three types of data to consider. In addition to structured type managed by a relational database management system, there are two other new types of data, including unstructured and semi-structured ones which are handled by the big data technologies. First, the structured data are those whose set of possible values is determined and known in advance. They respect a predefined model that allows them to be accessed and managed very easily. Structured data is often managed by relational database management systems in the form of tables, and it is handled using a query language such as SQL or PL/SQL. Secondly, the type is the unstructured data which is the opposite of the first type (structured). These are data that do not respect a data schema and are not organized in a predefined way. This type of data can have various formats such as videos, pdf, images, doc, text files, activities on social networks, etc. They are both complex and bulky, and traditional databases cannot manipulate or query them. In addition to structured and unstructured data, there is also a third category, called semi-structured data. Semi-structured data is information that is not stored in a structured dataset, but its structure contains tags that make it easier to manipulate and analyze. Examples of semi-structured data may include XML files, text emails, and JSON documents.

3.4 Big data and cloud computing

Big data technologies require processing power, speed of execution and huge storage space. This requires big computers with processor, memory, and disk space resources that offer tremendous computing power and performance. In this context, the services of cloud computing can be used. In fact, cloud computing and big data are two inseparable elements. Cloud computing offers theoretically infinite processing power and storage capacity, in addition to the availability of resources. Indeed, we cannot discuss the integration of big data and e-learning platforms without considering the cloud, because it gives the resources needed to deploy big data technologies and tools, as well as learning management systems (LMS). In addition, the cloud offers a preconfigured, ready-to-use environment that incorporates massive data processing technologies.

4. Integration of big data and online learning systems

4.1 Proposed approach for integrating big data, online learning systems, and cloud computing

Big data and the cloud have become key components of any information system, including e-learning systems. As a result, their integration is a major necessity to free themselves from hardware and technical architecture installation issues and to take advantage of the important volumes of data generated by such a system, as well as to gain flexibility in processing and analysis. Identifying useful information from learning data is a big challenge, especially with the significant increase in the amount of data produced every day by online learning platforms. To overcome this problem, big data ecosystem provides advanced technologies, methods, and techniques trough machine learning algorithms in the form of software libraries (APIs,

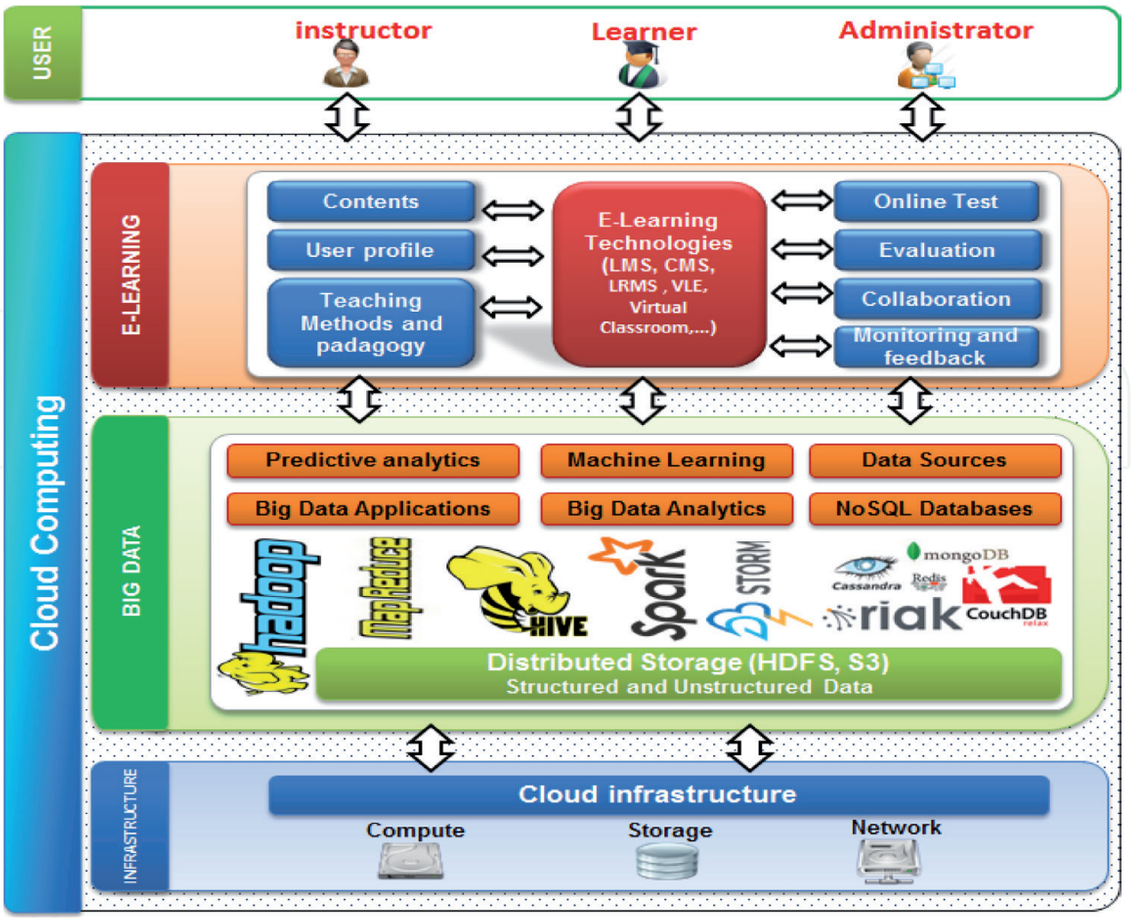

Figure 2.
Integration of big data, online learning systems, and cloud computing.

Frameworks, etc.) that are very powerful and easy to use. Such technologies make it possible to prepare and analyze, in a distributed manner, large amounts of data in order to make the best decision and to help e-learning professionals to be able to continuously enrich and enhance their strategies to be adapted to the interests and preferences of each learner [13] (**Figure 2**).

Infrastructure is the first layer that is the lowest level of the proposed approach. The infrastructure layer is built with compute, storage, and network resources that are virtualized and delivered as services through the cloud. The latter is responsible for providing virtual computing resources and the big data technologies needed to provide e-learning systems with a favorable execution environment. The resources of this layer are scalable. If, for example, when analyzing a massive volume of data, an application requires a huge amount of computing time or disk space, the cloud infrastructure will automatically expand to allocate the resources required by this application. This mechanism allows great flexibility over the traditional approach based on traditional hosting technologies in which server resources are limited. The great advantage of this layer is that it offers a scalable, resilient and fault-tolerant infrastructure.

The second layer of this approach is that of the big data ecosystem. The big data layer includes decentralized storage technologies and distributed large data, massively parallel computing, advanced analysis, optimization and visualization of the processing results. It groups together various big data technologies which can be classified in:

- Distributed file systems: stores data in multiple nodes in a cluster in a replicated manner to provide redundancy and high availability. HDFS remains best known as an open source solution for distributed data storage and management.

- NoSQL movement: represent the new generation of databases management systems. They allows moving away from the relational model and overcoming the limitations of RDBMS in terms of the amount and types and formats of the data handled. Among the distributed databases are CouchDB (document-oriented), Cassandra and Hbase (column-oriented), which do not impose strict schema rules as in the case of the relational model.

- Distributed processing and predictive analysis infrastructure: allows parallel computing of large data sets across machine clusters. The best known open source example of distributed systems is Hadoop developed by the Apache Foundation. They are also tools implementing mathematical methods applied to computing for the analysis of giant databases via predictive models of machine learning. In addition, Apache Spark is a high-performance framework dedicated to the design and creation of large-scale applications for predictive analysis in memory.

The e-learning system represents the third level. This is the application layer represented by the e-learning system containing, in particular, learning management tools (LMS), content management systems (CMS), virtual learning environments (VLE), etc. The information in this layer can be data in the form of educational content, information on the learner or teacher profile, course registrations, etc. These data play a very important role and will be useful in generating personalized learning resources by adapting learning content to the needs of each learner to provide a more appropriate learning platform. To perform this adaptation mechanism, the e-learning system must use the technologies of the lower layer (big data) to exploit advanced predictive models by applying parallel algorithms of machine learning on the learning data.

Generally, an online learning is a platform consisting of hardware, software, and user. The hardware includes memory capacity, network bandwidth, and CPU provided by the IaaS services of cloud. The e-learning system is the software. Users are actors who use the system to communicate, store and process information. The users are mainly the learner, the teacher, and the system administrator. In addition, a fourth player is data scientist can be added who is responsible for configuring, installing, monitoring and controlling the distributed environment of the cluster, as well as developing and deploying analytics models and implementing data mining techniques on platforms. Big data.

4.2 Methodology

Generally, in the different areas, it is a priority to have a clear methodology before starting the implementation phase or the operationalization of a project. This methodology shows the process and the mechanism to better manage big data projects. So, we must address a personalized methodology adapted to the context of e-learning. To do this, we identify the key steps in the big data process that, from the sources of data generated by online learning platforms, extract value and insights to help educational distance to make good decisions. This includes the acquisition, discovery, preparation, modeling, processing and visualization of the results of the data analysis. When data is effectively captured, processed and analyzed, e-learning professionals can have a complete understanding of their learners, educational resources, assessment results, and so on. In this way, they will be able to offer more personalized learning activities, produce relevant teaching strategies, provide adaptive learning to each learner, and improve the quality of educational content. **Figure 3** schematizes our methodology for using big data in

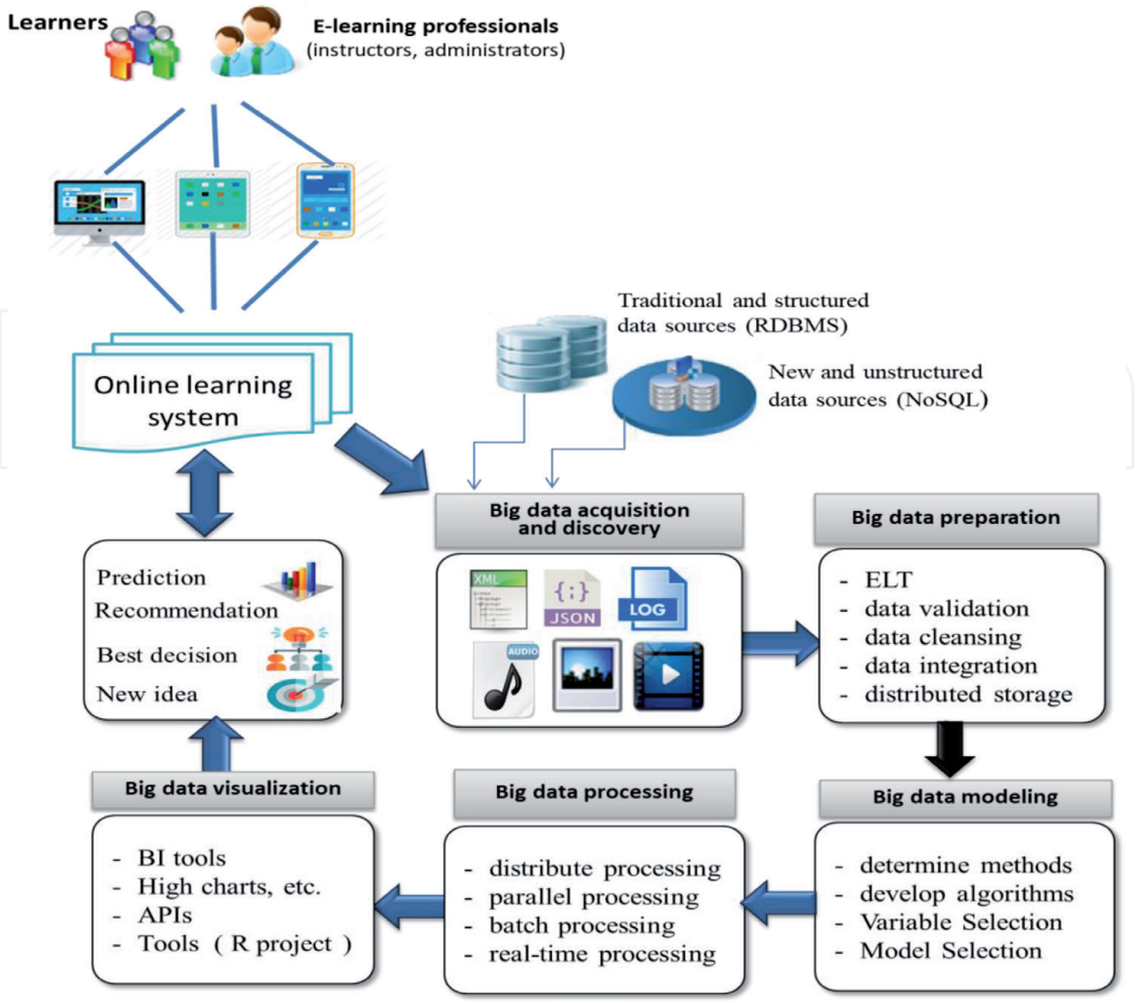

Figure 3.
Methodology for dealing with massive data in online learning.

distance learning environments. It describes the different stages, including big data acquisition and discovery, preparation, modeling, processing, and visualization.

Actually, in order to handle the large volume of data produced by online learning platforms, the data itself must go through a series of five steps:

discovery and acquisition phase describes the process of collecting and discovering the data produced by the distance learning environments, which can be information about the learner (profile, knowledge, skills, etc.), and educational resources including all formats (text, image, video, web page, etc.). They also can captured from learners' interactions between learners and teachers through social networks, wikis, and forums.

The second phase of our methodology is that of data preparation that comes just after the acquisition and discovery phase. It is a coherent set of operations that retrieve, load, and transform (ELT) multiple data sources. Indeed, this phase includes the integration of the data generated by the e-learning platforms, prepare them and transfer them to be stored in a distributed file system or a NoSQL database such as Cassandra, HBase, etc. Data preparation is a crucial step in the analysis process because it is at this level that we have to filter the collected data to rule out unnecessary or noisy data such as redundancy and keep only those that are relevant and good quality, something that will be used later as input of the analytical model.

Big data modeling is the third phase of our methodology. It aims to specify a suitable method to take advantage of large datasets. To do this, it is necessary to determine the right model to apply on these data. Indeed, during this phase, we

must identify the model, determine the appropriate method to use and develop the appropriate algorithm to implement the chosen method. In this sense, big data technologies implement various large-scale machine learning techniques, including: classification, clustering, association rules, regression, collaborative filtering. The best model depends on the type, quality, and size of the data to be analyzed and the available computing resources. It also involves data explorion to learn more about the relationships between variables and then selecting the most appropriate key variables for such a model.

Big data processing represents the fourth phase of our methodology. Actually, big data relies on a parallel computing approach to deal with the intensive processing needs and the increase in data volume. Technologies dedicated to data manipulation are transformed chronologically into batch processing, real-time processing and hybrid computing. Batch processing try to solve to the volume problem, real-time computing respond to speed issues, and hybrid computing is good for both. Distributed computing systems are widely developed, primarily to support the analysis of gigantic data, including Frameworks Hadoop MapReduce [14] and Apache Spark [15].

This phase aims to clearly and effectively visualize and communicate the results of the analysis of the learning data through rich tools and advanced software libraries to synthesize the information of the treatment. This often contains tables, in the form of graphical representations, such as curves, bars, sectors, and histograms. It also describes a set of techniques, software and utilities designed to help e-learning professionals to have a clear view of the enormous data generated by learners.

5. Implementation and results

To validate our model based on the integration of big data technologies and methods in the context of CEHL, a course recommender system has been implemented to show the effectiveness and usefulness of our contribution. It is about a recommendation engine acting as a predictive unit able to adapt to a given learning profile by anticipating its next actions through the suggestion of relevant pedagogical resources that best meet its preferences and interests. The aim is, therefore, to provide the learner with a personalized educational and pedagogical plan according to his profile. For this purpose, our recommender system uses advanced machine learning techniques, especially association rules method for extracting knowledge through the analysis of learning traces. The realization of this system was made using the historical data of learner's activities collected from the ESTenLigne platform of the Higher School of Technology of Fez. Our recommender system relies on a totally distributed architecture which consists in setting up a large-scale course recommender system. It was developed and tested using the FP-growth parallel algorithm, and the Apache Spark Framework. The deployment of this version is done through the Hadoop distributed cluster infrastructure.

5.1 ESTenLigne project

The present work is a part of the ESTenLigne [16] project, which is the result of several years of experience for the development of e-learning in the Sidi Mohamed Ben Abdellah University of Fez. It was started since 2012 by the EST network of Morocco, which aims the development of distance education based on new information and communication technologies through the implementation of open, adapted and free online learning platform, and taking into account the dimensions of exchange, sharing and mutualization of pedagogical resources [17, 18]. Several

works have been done as part of this project including the training of experts across e-learning in the context of the Coselearn I project, and teacher training through Franco-Moroccan EST [19] and IUT [20] cooperation [18, 21]. Furthermore, there are some researches that have been done around this project such as the analysis of the use of educational resources where the objective was to analyze the use of pedagogical resources in some courses namely the algorithmic course [22]. Also, a case study for collaboration analysis of online course based on activity theory [23]. In addition, the development an e-learning recommender system based on R environment [24].

In fact, the students have a lot of difficulties and are lost in the diversity of educational resources, particularly the large number of available courses. This requires the adaptation of the teaching to meet the needs of students. To solve these problems, we develop a course recommender system to promote learning to learners through creating a smart solution. It is able to generate the most appropriate courses automatically based on historical data of learner's activities.

5.2 Association rules method

In the area of machine learning, Association rules [25] is an unsupervised learning method, widely used in many areas, including referral engines, online purchase transaction analysis, and flow analysis. Clicks on multiple web pages [26]. Its purpose is to discover relationships between variables in a set of data, which we will call transactions, in the form of interesting association rules. In other words, this method consists in detecting associations between data stored in a giant database. It is a set of powerful exploratory techniques widely used in many sectors but also for scientific research purposes. The most popular application using the association rules is the one concerning the analysis of consumption habits. The power of the association rule method lies in its ability to extract hidden structures in a massive amount of data.

Generally, association rule technique produces a large number of rules, but to select interesting rules in the set of generated relationships, it has two important criteria for determining the quality of a rule by measuring its strength, namely: the minimum thresholds of support and confidence. Support is the percentage (%) of transactions containing the set of items X, while confidence is defined as the percentage (%) of transactions containing X, which also contain Y. Therefore, a force association rule $X \Longrightarrow Y$ should satisfy: supp $(X \cup Y) \geq \sigma$ and conf $(X \Longrightarrow Y) \geq \delta$, where σ and δ represent the minimum threshold of support and confidence, respectively.

As part of our research, we applied the rules of association technique in the context of a computer environment for human learning dedicated to e-learning. Therefore, a transaction in our case is represented by the learner profile. Likewise, the items are replaced by all available resources in the database. A transaction is represented by a learner's enrollment in a number of courses during his or her learning path. We can therefore define the support (1) and the confidence (2) as follows:

$$\text{supp}(X \Longrightarrow Y) = \frac{\text{nombre d'apprenants inscrits aux X et Y}}{\text{nombre total d'inscriptions d'apprenants}} \quad (1)$$

$$\text{conf}(X \Longrightarrow Y) = \frac{\text{nombre d'apprenants inscrits aux X et Y}}{\text{nombre d'apprenants inscrits aux X}} \quad (2)$$

5.3 FP-growth algorithm

FP-growth [27] (Frequent Pattern Growth) is a very powerful algorithm for extracting the most frequent elements from large data sets by allowing a very fast discovery of the association rules without generation of candidates, which requires

more memory and time processor. In fact, the generation and testing of candidates requires several analyzes of the database. By using FP-growth, the number of database scans is reduced to two. The first scan aims to count the support of each item, the non-frequent items are deleted, while the frequent items are sorted in descending order of support, in the form of a list of frequent items (L). Then, in the second scan the algorithm builds the FP-tree structure with the creation and insertion of the different nodes. These operations constitute the first step of the algorithm. On the other hand, the second step is to extract sets of frequent elements from the constructed FP tree. The FP-growth algorithm is based on the "divide and conquer" strategy of breaking down a problem into subproblems. First, it compresses frequent itemsets represented in the database using a compact data structure called FP-Tree (frequent-pattern tree) whose branches contain the possible item associations. The FP-growth method transforms the problem of finding the longest frequent itemset by searching for the smaller one and its concatenation with the corresponding suffix. This reduces the cost of research.

5.4 Big data technologies and its components

This section introduces all the big data technologies used as well as their different components. Each technology has a definite role and participates in the process of extracting prediction knowledge by providing a list of highly recommended courses according to the needs and interests of each learner. These technologies are deployed on a cluster infrastructure of nodes that are interconnected via network protocols to be able to communicate and exchange data during the analysis of learning traces. These technologies can be organized in three layers. **Figure 4** describes the different big data frameworks used to develop the large-scale course recommendation system.

All implemented technologies can be grouped in 3 layers:

First, there is the layer of distributed data storage. We chose to use Hadoop's Distributed File System (HDFS). In fact, HDFS is a fault-tolerant file system capable

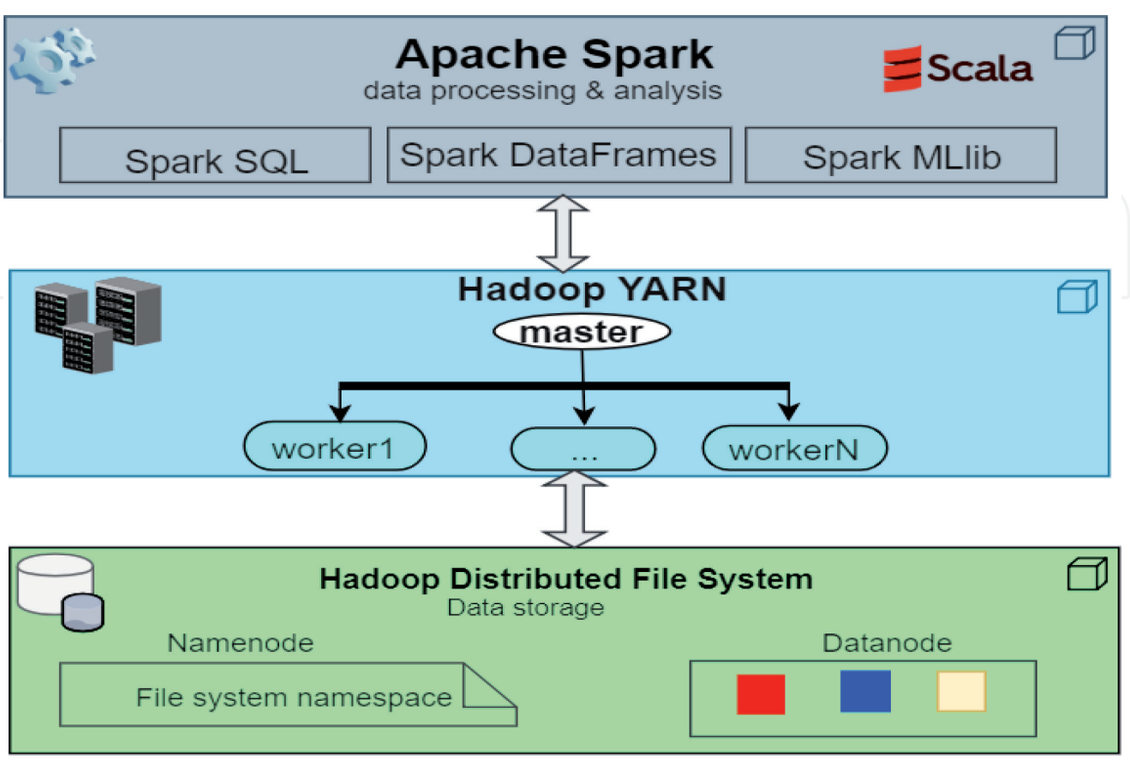

Figure 4.
Big data technologies.

of managing distributed data across large clusters. It has a master/worker architecture. HDFS provides high performance access to large amounts of data. It creates an abstraction of hard drive resources to allow the management of distributed physical storage of multiple nodes as if there is only one storage space. In the HDFS architecture, data is managed across the cluster, in different Datanodes, by the workers in the form of block-structured files. The locations of these blocks and the namespace of the files and directories are kept in the Namenode component in the master node [28].

In the second level, Yarn [29] is found as the node cluster resource manager. This is a Hadoop module dedicated to scheduling and executing tasks, at the same time, on a number of computers in a cluster. It is also responsible for managing disk resources, memory, CPU and cluster network. Yarn's main idea is to separate resource management from the computational model. Indeed, Yarn will take care to rent the necessary resources and to distribute the basic tasks on different units of calculation of the machines of a cluster.

Finally, the upper layer represented by the Spark Framework [15] is responsible for the manipulation and analysis of the data. It is used to apply association rules techniques to learner learning data, collated from the ESTenLigne project. To implement our course recommendation system, Scala was chosen as the development language. The advantage of Spark is its ability to support multiple programming languages such as Java, Python and R. This framework provides many libraries. In our use case, we focused on just three components: Spark SQL, Spark DataFrames, and Spark MLlib.

- Spark SQL allows you to connect to the moodle LMS database and execute SQL queries.

- Spark DataFrames is a Spark module for structured data processing.

- Spark MLlib implements several machine learning algorithms, including the FP-growth parallel algorithm.

5.5 Distributed system architecture

In general, our approach is to generate recommendations by analyzing the traces of learning activities that are the source of knowledge in the process of personalization of learning resources provided to learners. The entry of the system thus consists of the history of course registrations, imported from the database of the e-learning platform. **Figure 5** describes this architecture in detail.

In the beginning, we need to load the data produced by the learners' interactions with the ESTenLigne platform. Then, this data, loaded by the Spark SQL library, is processed in a distributed manner using the Spark Framework that runs on a Hadoop cluster and uses the Yarn Resource Manager. Indeed, Apache Spark provides a special library dedicated to machine learning techniques, called MLlib. This library proposes an implementation of the parallel FP-growth algorithm in the Scala language. Subsequently, the prepared data is analyzed using the FP-growth algorithm of the Spark MLlib library. Then, Spark connects to Hadoop HDFS to store the data on machine clusters. Then, the recommendation system generates the catalog of the most relevant courses. Finally, the results of the recommendation engine can be presented to the user in order to guide them and suggest the educational resources most suited to their interests. Thus, the learner can go through the courses recommended by our system and begin to learn those that suit his cognitive level and preferences.

Besides, our system uses the open source tool Ganglia. It is a highly scalable, distributed and scalable solution for monitoring large environments such as clusters

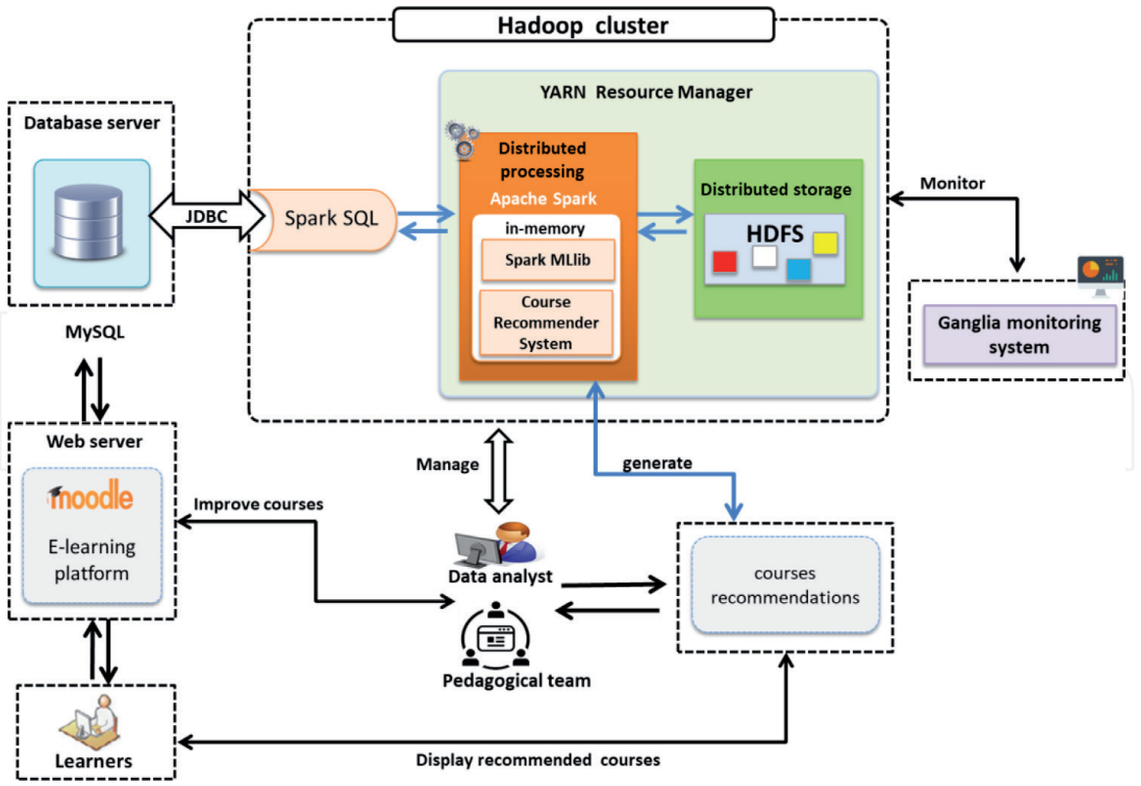

Figure 5.
Distributed architecture of course recommendation system.

and grids, as well as measuring performance and resource consumption such as CPU utilization, memory, and data storage of each node of the cluster. It also controls and visualizes network traffic, such as bandwidth usage or the amount of data transported over the network.

5.6 Setting up the big data environment

Considering the high cost of installing a physical infrastructure of big data clusters, the use of virtualization tools for a distributed architecture represents an alternative solution to configure, develop, test and validate our course recommendation system. The Hadoop multi-node configuration is done in a fully distributed environment consisting of three nodes. The machine with the IP address: 192.168.56.101 works as the Hadoop master, which contains the Hadoop components, namely ResourceManager and NameNode. The master node runs Hadoop processes to manage and coordinate cluster tasks and services. In fact, the virtual machine identified by the IP address 192.168.56.101 works as a master and worker at the same time. The other machines in the cluster are workers. The worker nodes are responsible for running the processes or the basic tasks of the parallel application. They also provide resources to the cluster to perform the processing of data assigned by the master. As shown in **Figure 6**, they respectively have machines with the IP addresses 192.168.56.102 and 192.168.56.103.

In order to build our big data infrastructure, we have prepared a cluster of three virtual machines. As a virtualization solution we used the free and popular VirtualBox solution. With the available computing resources, we created three virtual machines by installing the Ubuntu 18.04.1 LTS operating system on each node of the Hadoop cluster. These nodes are connected to each other using a private LAN. The capacity and configuration of all virtual machines are described in **Table 1**.

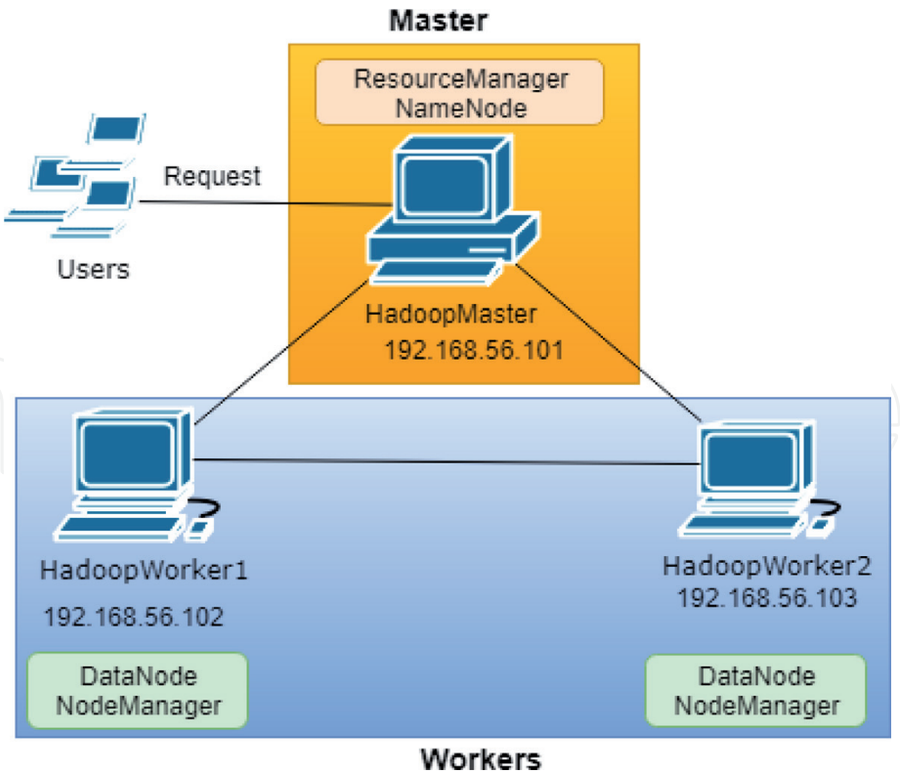

Figure 6.
Hadoop cluster configuration.

Machine	Network	Cores	Memory	Disk
Master	192.168.56.101 (master.domain.com)	8 Core i5 (7th Gen)	8 GB	32 GB
Worker1	192.168.56.102 (worker1.domain.com)	8 Core i5 (7th Gen)	8 GB	32 GB
Worker2	192.168.56.103 (worker2.domain.com)	8 Core i5 (7th Gen)	8 GB	32 GB

Table 1.
Configuration of Hadoop cluster nodes.

After configuring the virtual cluster network, we unzipped and installed Hadoop 3.1.1 and Spark 2.3.1 at the master node of the cluster. Then we moved the installation folder of both Frameworks to the worker nodes using the SSH protocol. Similarly, we used the power of Secure Copy (SCP) to get the same copy of Apache Hadoop and Spark. The Java version 1.8 has been installed on each node and we have configured a password-free ssh between the nodes so that the master Hadoop node can connect, start, stop, and execute tasks in different workers.

6. Results

To run parallel FP-growth, we had to specify the minimum support and confidence thresholds in order to find the strongest correlations between course enrollments in learning activity traces. The number of interesting association rules changes according to the value of the support, the confidence and the size of the database. Thus, we used the minimum support threshold of 5% and we set 60% as the minimum confidence threshold. In fact, the course recommendation system generates two types of results that meet the specified support and confidence criteria. First, it finds the list of frequent courses in the database of the e-learning system based on the calculation of the support and the confidence of each itemset, it keeps only the list of itemsets that

satisfy the condition of the minimum threshold of confidence. The 10 main rules of interest, ordered according to the confidence measure, are shown in **Table 2**.

According to the results obtained in **Table 2**, the rule of association between courts {11 and 46} and {45} has the greatest confidence. Similarly, the rule of association between the courts {7} and {6} has the lowest confidence. Based on calculated values and support and confidence, we can clearly identify the courses most likely to be followed by learners, and most relevant to recommend them. For example, rule 1 {11, 46} = > {45} has the greatest confidence, that is, the strongest. Our system therefore suggests course 45 to students who have already enrolled in courses {11 and 46}. According to the results of **Table 2**, the confidence of rule 1 is 100%, because, at the level of the database, we find that 56 students have enrolled in courses {11 and 46}, and that these 56 them also enrolled in course {45}. For association rule number 2, the analysis of historical student registration data from the ESTenligne platform shows that 57 students took courses {11 and 45}, and 56 of them followed also the course {46}. So, the confidence of association rule 2 is 98%. Therefore, our recommendation system recommends the course {46} to students enrolled in courses {11 and 45}. With regard to association rule number 3, there are 123 learners enrolled in the course {46}, of which 118 are also enrolled in course {45}. Thus, the confidence of rule 3 is 95%. Our system therefore offers the {45} course to students enrolled in {46}. For the 10 strongest association rules, we note that the confidence values are between 0.69 (69%) and 1.00 (100%), which demonstrates that we have achieved good results. We can therefore conclude that the course recommendation system that we have proposed can provide relevant teaching resources by recommending courses suitable to each learner in order to guide them through their learning path.

After establishing the recommendation model using the training dataset (70%), it can also be used to predict the result of recommending courses to learners through the database. Test (test dataset) (30%). The recommendation system generates a catalog of courses for each learner profile. **Table 3** shows the top 10 predictions out of a total of 48 predictions.

The results of the prediction of the recommendation system that we have developed gives for each learner (id) the list of courses in which he participates and a catalog of the predictions of the courses. For example, our course recommendation system suggests courses 18 and 7 to the learner (id = 541) who is already enrolled in courses 46, 6, 11 and 45. Also, he recommends course 14 to the learner (id = 720) which follows courses 45, 46 and 15 courses, etc.

Rule	Antecedent	Consequent	Confidence
[1]	[46, 11]	[45]	1.000
[2]	[45, 11]	[46]	0.975
[3]	[46]	[45]	0.952
[4]	[45]	[46]	0.919
[5]	[7, 6]	[18]	0.868
[6]	[7]	[18]	0.818
[7]	[6, 18]	[7]	0.785
[8]	[7, 18]	[6]	0.733
[9]	[6]	[18]	0.711
[10]	[7]	[6]	0.690

Table 2.
Parallel FP-growth results.

Learner ID	Items (course)	Recommendations
541	[46, 6, 11, 45]	[18, 7]
720	[45, 46, 15]	[14]
19	[6]	[18, 7]
277	[18, 14, 6, 9, 3, ...	[43]
287	[17, 43, 42, 9, 2, ...	[6]
155	[15, 17, 42, 9]	[14, 43]
1157	[6]	[18, 7]
184	[40, 6, 43, 42]	[18, 7]
274	[7, 6, 15, 4, 3, ...	[43]
766	[15, 45, 46]	[14]

Table 3.
Prediction results of the course recommendation system.

7. Conclusion

This chapter aims to integrate the new generation of information and communication technologies, especially the big data ecosystem, in Computing Environments For Human Learning dedicated to online learning. In fact, big data provides a wide range of tools and systems for distributed storage, massively parallel processing, and predictive analytics. This set of technologies can be used in the processing and analysis of massive data produced by learners interactions. It offers high-level frameworks allowing a lot of advantages to greatly improve the quality and disponibility of distance learning platforms.

So, our model can really improve the online learning field which every learner can have the maximum benefits from that. Furthermore, pedagogical teams and administrators of e-learning platforms have valuable tools and advanced APIs for analyzing data in order to improve learning strategies, make better decisions and offer a big variety of new learning methods.

To implement big data technologies in e-learning systems, this chapter has designed and developed a course recommender system to provide an adaptive learning solution, which consists of adapting teaching resources to individual preferences and needs of each learner. The implementation of the proposed course recommendation engine uses machine learning techniques, in particular, the association rules technique to find all the interesting relationships from historical student enrollment data. The results obtained show the effectiveness of our system in terms of quality and relevance of course recommendation and execution time performance thanks to the decentralization approach of the processing and analysis of data. The deployment of our system is done in the distributed infrastructure of Hadoop and the powerful in-memory processing and advanced analysis of the Spark Framework. The implemented Spark application is based on a completely different approach which consists, in fact, in distributing the processing and the data on several nodes each of which carries out specific tasks in order to execute the FP-growth algorithm in parallel on different machines of the cluster. The distribution mechanism of computing and storage solves the problem of limiting available resources while speeding up the execution speed and the cost of processing.

Author details

Karim Dahdouh*, Ahmed Dakkak, Lahcen Oughdir and Abdelali Ibriz
Engineering Sciences Laboratory, FPT, Sidi Mohamed Ben Abdellah University, Taza, Morocco

*Address all correspondence to: karim.dahdoh@gmail.com

References

[1] Bruillard E. Les Machines à Enseigner. Paris: Hermès; 1997

[2] Guri-Rosenblit S, Gros B. E-learning: Confusing terminology, research gaps and inherent challenges. International Journal of E-Learning & Distance Education/Revue Internationale Du e-Learning et La Formation à Distance. 2011;**25**(1). Available from: http://www.ijede.ca/index.php/jde/article/view/729

[3] Udupi PK, Malali P, Noronha H. Big data integration for transition from E-learning to smart learning framework. In: 2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC). Muscat, Oman: IEEE; 2016. pp. 1-4. DOI: 10.1109/ICBDSC.2016.7460379

[4] Birjali M, Beni-Hssane A, Erritali M. Learning with big data technology: The future of education. In: Abraham A, Haqiq A, Hassanien AE, Snasel V, Alimi AM, editors. Proceedings of the Third International Afro-European Conference for Industrial Advancement — AECIA 2016. Vol. 565. Cham: Springer International Publishing; 2018. pp. 209-217. DOI: 10.1007/978-3-319-60834-1_22

[5] Logica B, Magdalena R. Using big data in the academic environment. Procedia Economics and Finance. 2015;**33**:277-286. DOI: 10.1016/S2212-5671(15)01712-8

[6] Cox M, Ellsworth D. Application-controlled demand paging for out-of-core visualization. In: Proceedings. Visualization '97 (Cat. No. 97CB36155). Phoenix, AZ, USA: IEEE; 1997. pp. 235-244. DOI: 10.1109/VISUAL.1997.663888

[7] Ghemawat S, Gobioff H, Leung S-T. The Google File System. ACM Édition; 2003. Available from: https://static.googleusercontent.com/media/research.google.com/fr//archive/gfs-sosp2003.pdf

[8] Hadoop. Apache Hadoop. Apache Software Foundation; 2019. Available from: http://hadoop.apache.org/

[9] Chakhari A. La digitalisation est une guerre mondiale armez-vous. 2015. p. 70

[10] Gartner. Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute; 2011

[11] NIST Big Data Public Working Group Definitions and Taxonomies Subgroup. NIST Big Data Interoperability Framework. Definitions. NIST SP 1500-1. Vol. 1. National Institute of Standards and Technology; 2015. DOI: 10.6028/NIST.SP.1500-1

[12] Grandmontagne Y. Les 5 V du Big Data—IT Social | Média des Enjeux IT & Business. Innovation et Leadership. 2014. Available from: https://itsocial.fr/articles-decideurs/les-5-v-du-big-data/

[13] Dahdouh K, Dakkak A, Oughdir L, Messaoudi F. Big data for online learning systems. Education and Information Technologies. 2018;**23**(6):2783-2800. DOI: 10.1007/s10639-018-9741-3

[14] Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. Communications of the ACM. 2008;**51**(1):107

[15] Apache Spark. 2019. Available from: https://spark.apache.org/

[16] ESTenLigne Platform. 2018. http://elearn.est-usmba.ac.ma/

[17] Ibriz A. Une Démarche Innovante de Conduite de Projet Elearning: C.D.I.O. In: 2ème Congrès International du Génie Industriel et du Management Des Systèmes (CIGIMS). 2015

[18] Ibriz A, Abdellatif SAFOUANE. L'Innovation Pédagogique dans les EST du Maroc: Le model et la Conduite d'un cas réussi à travers le Projet ESTenLigne. In: Colloque Eomed. 2014

[19] Ecole Supérieure de Technologie. 2018. Available from: http://www.est-usmba.ac.ma/

[20] Instituts Universitaires de Technologie. 2018. Available from: http://www.iut.fr/

[21] Oughdir L, Ibriz A, Harti M. Modélisation de l'apprenant dans le cadre d'un environnement d'apprentissage en ligne. In: TELECO2011 & 7ème JFMMA Mars 16-18, 2011—Tanger MAROC. 2011

[22] Benslimane M, Kamar O, Mehdi T, Mohammed B. Proposal of an approach of online course design and implementation: A case study of an algorithmic course. 2016;7:7

[23] Ibriz A, Benslimane M, Ouazzani K. Didactics in online learning technical courses: A case study based on activity theory. International Journal of Computer Science and Information Technologies. 2016;7:849-854

[24] Dahdouh K, Oughdir L, Dakkak A, Ibriz e A. Smart courses recommender system for online learning platform. In: 2018 IEEE 5th International Congress on Information Science and Technology (CiSt). Marrakech: IEEE; 2018. pp. 328-333. DOI: 10.1109/CIST.2018.8596516

[25] Rakesh A, Srikant R. Fast algorithms for mining association rules. In: Proceedings of the 20th VLDB Conference Santiago, Chile; 1994. p. 13

[26] Larose DT. Data Mining and Predictive Analytics. Hoboken, New Jersey: Published by John Wiley & Sons, Inc.; 2015. p. 827

[27] Li H, Yi W, Zhang D, Zhang M, Chang e EY. Pfp: Parallel Fp-growth for query recommendation. In: Proceedings of the 2008 ACM Conference on Recommender Systems—RecSys '08. Lausanne, Switzerland: ACM Press; 2008. p. 107. DOI: 10.1145/1454008.1454027

[28] The Hadoop Distributed File System (HDFS). 2018. Available from: https://hadoop.apache.org/docs/r3.1.1/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html

[29] Apache Hadoop YARN. 2019. Available from: https://hadoop.apache.org/docs/r3.1.1/hadoop-yarn/hadoop-yarn-site/YARN.html