Big Data, the Next Step in the Evolution of Educational Data Analysis

W. Villegas-Ch[1(✉)], Sergio Luján-Mora[2], Diego Buenaño-Fernandez[1], and X. Palacios-Pacheco[3]

[1] Facultad de Ingeniería y Ciencias Agropecuarias, Universidad de Las Américas, Quito, Ecuador
{william.villegas,diego.buenano}@udla.edu.ec
[2] Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Alicante, Spain
sergio.lujan@ua.es
[3] Departamento de Sistemas, Universidad Internacional del Ecuador, Quito, Ecuador
xpalacio@uide.edu.ec

Abstract. This paper presents an analysis of new concepts such as big data, smart data and a data lake. It is to sought integrate learning management systems with these platforms and contribute to education by making it personalised and of quality. For this study, the data and needs of a university in Ecuador have been considered. This university has set its goals to the discovery of patterns, using data mining techniques applied to cubes generated in a data warehouse. However, the institution wants to integrate all the systems and sensors that contribute to the educational development of the student. Integrating more systems into the data warehouse has compromised the veracity of the data and the processing capabilities have been surpassed by the volume of data. The paper proposes the use of one of the platforms analysed and its tools to generate knowledge and to help the students to learn.

Keywords: Analysis of data · Big data · Data lake · Data mining
Data warehouse · Smart data

1 Introduction

Education currently uses learning platforms, information and communications technology (ICT) to manage learning. The aim of this integration of pedagogy and ICT is to create learning methods that are accessible and used by students. The integration uses, as its main tool, the learning management system (LMS) [1]. LMSs have become the main repository of student performance information. In order to take advantage of this information, data mining techniques are used to obtain patterns in student performance [17]. For this work the Moodle platform of a university in Ecuador is used. In this university, the use of Moodle has been institutionalised and policies of use have been created for a standard management of the courses of each one of the teachers. The Moodle platform has been customised according to the policies and needs presented by

Á. Rocha and T. Guarda (eds.), *Proceedings of the International Conference on Information Technology & Systems (ICITS 2018)*, Advances in Intelligent Systems and Computing 721,
https://doi.org/10.1007/978-3-319-73450-7_14

the academic department in charge of evaluating student learning [7]. The customisation of the platform is based on the integration of modules and links that allow the proper management of academic resources which students can use for the development of activities. The large amount of data generated has surpassed the analysis capabilities to be processed in a conventional way. To supplement this processing, it is necessary to work with new concepts such as big data, smart data, data lake and data mining. The convergence of these concepts with the educational systems will allow evaluation of the data and transform it into useful information. This paper defines which of the platforms analysed, based on the needs of an educational institution, gives greater benefits for decision making and contributes to the improvement of education.

The work is composed as follows: Sect. 2 presents the concepts used in the development of this work; Sect. 3 determines the method used for the analysis of big data, smart data and the data lake; Sect. 4 presents the analysis of results to choose the platform that gives the solution to the problems raised, and Sect. 5 presents the conclusions that have been reached from the work done.

2 Preliminary Concepts

This article takes into account several key concepts that help the management and application of different analyses that help clarify new processes within the educational field.

2.1 Big Data

Big data is born from data sets or combinations of data sets whose size, complexity (variability) and speed of growth (velocity) make it difficult to capture, manage, process or analyse using conventional technologies and tools [10]. Big data requires a combination of different tools such as relational and non-relational databases [5]. It makes use of analytical tools to transform data into value information. Big data analysis is done on hundreds or thousands of blade servers. The tasks are distributed in intelligent networks of parallel processing that allow the use of analyses in a real time to customise, segment, optimise prices and relate in with customers [6].

Big data allows the management and analysis of huge volumes of data that cannot be processed in a conventional way. The size used to determine if a dataset is considered big data is not defined and keeps changing over time [9]. However, as a benchmark, analysts and professionals refer to datasets ranging from 30–50 TB to several petabytes.

Big data is a complex data set; this is mainly due to the unstructured nature of much of the data generated by modern technologies. These technologies are: web logs, radio frequency identification and built-in sensors in devices, vehicles, Internet searches, social networks, smart phones, GPS devices and call centre records.

Organizations have handled large volumes of data for a long time and have developed data warehouses and powerful analytical tools. These tools allow the adequate handling of large volumes of data [2]. The goal of big data is to turn the data into information that facilitates decision making in real time. However, more than a matter

of size, it is a business opportunity. Organizations use big data to understand the profile, needs and feelings of their customers regarding the products or services offered. In order to use big data effectively, it must integrate structured data from a conventional business application, such as enterprise resource planning (ERP) or customer relationship management (CRM).

2.2 Smart Data

The concept smart data appears following the big data and focuses on processing data, in order to convert them into statistics. These statistics serve to find the content of higher value (separates the useful from the useless), smart data is a complex data filter [5]; it is a concept that revolves around mass information management but only of the one that has a real value.

To understand the smart data easily, the big data would be information gathering, processing and filtering. The smart data would act once all that processed information is available and use mathematical formulae to convert data into "axiomatic" responses on a market [4].

2.3 Data Lake

A data lake is a storage repository that contains a large amount of raw data and is kept there until needed [8]. A data lake is a concept close to data warehouse that allows for the storage and processing of large volumes of data. They are used to collect raw data before the data sets pass into a production analytical environment, such as a data warehouse [11].

The main benefit of a data lake is the centralisation of disparate content sources. Once assembled, these sources can be combined and processed using big data (searches and analyses). The disparate content sources often contain confidential information that will require the implementation of appropriate security measures in the data lake.

The content of the data lake can be normalized and enriched. This may include extracting metadata, format conversion, augmentation, entity extraction, crosslinking, aggregation, de-normalization or indexing. Scattered users around the world can have flexible access to a data lake and its content from anywhere. Accessibility increases re-use of the content and helps the organisation gather the data needed more easily to drive business decisions.

2.4 Data Mining

Data mining is the analysis stage of knowledge discovery in databases (KDD) [3]. It is a field of statistics and computer science that attempts to uncover patterns in large volumes of data. It uses the methods of artificial intelligence, machine learning, statistics and database systems. The general objective of the data mining process is to extract information from a set of data and transform it into useful information. It finds repetitive patterns, trends or rules that explain the behaviour of data in a given context. The data are the raw materials, the user attributes some special meaning to them and they become information; the specialists elaborate or manage a model, so that the

interpretation that arises between the information and that model represents an added value, which is called knowledge.

3 Method

For the development of this work, the analyses of big data, smart data and the data lake have been undertaken. The objective is to specify how these three concepts can be applied to educational institutions and which will help to improve education and the generation of knowledge. For this analysis, technical data such as data volume, historical management needs, pre-processing of data, etc. are considered.

Once the results are obtained, a method is proposed that helps to discover which of the platforms provides the best solution for the needs of the knowledge analysis within the LMS. The tool should detect patterns in student behaviour that help define strategies for resource improvement, as well as educational activities provided on LMS platforms.

3.1 Description of the Problem

For this work we analyse the data of a university in Ecuador. Since 2010, this educational institution has worked with the Moodle platform as an e-learning tool. The institution has eight thousand students in different programs. The students, to finish their programs, must culminate ten educational periods. Each period consists of four months and on average six subjects are given. The use of the Moodle platform is obligatory in support of the teaching-learning activity. For the control of the use of the platform, the administrators generate reports of the activities performed at the end of each period.

At the beginning of each educational period, a new virtual classroom is generated for each subject within the LMS. The generation of virtual classrooms has the purpose of removing the virtual classrooms from past periods and, in such a way, maintains backups and records of the data. With the passage of time, standards have improved the use of the platform, in the same way that modules have been created to help teacher management. These improvements have allowed the creation of new techniques and resources that help the development of the learning in the students. Currently the platform has links to different websites, multimedia material and games. It has even been integrated with tools that allow virtual tutorials, as well as systems that allow the detection of similarity between documents or any file that is on the internet.

The registry of activities in the platform has made it the most important repository that handles educational data within the institution. The constant growth of data volume has forced to implement a data warehouse for data analysis. The analysis is based on the conventional data mining application adapted from business analysis to data cubes. Data mining has revealed patterns in the behaviour of educational data and improved learning techniques. However, the data warehouse tool has fulfilled its life cycle and exceeded the processing capabilities compromising the accuracy of the results. To solve these problems, there needs to be an improvement in the design of the data

warehouse or, in turn, to apply one of the new trends in data analysis and to take advantage of the information that has been generated during this time.

For our work we consider the following data from an Ecuadorian university. For the calculation, the number of hours that an average student with 7/10 grades devotes to the Moodle platform has been considered, as Table 1 shows.

Table 1. Calculation of hours of use of the Moodle platform

# hours of an average student on platform per week	Days per week	# weeks per period	# of periods per year	Total hours per year
5.5	5	16	2	880
Total number of students			8000	7040000

Table 2 calculates the average storage utilization used by each student over the course of a year. These data are considered important for analysis because they give us an exact figure on the consumption that has been generated in the database of the Moodle platform. The calculation has been made with the known data that are: the 5 TB of storage used from 2010 to 2017 and the 8000 students with which it counts, the institution until the indicated year. The result is 0.62 GB for each student since 2010 then we have detailed the storage consumption of each student per year.

The amount of storage shown in Table 2 indicates that it is a very low growth. However, due to data analysis needs, academic authorities have proposed increasing the management of the LMS platform and generating data by 50% with respect to megabyte used per year and per student.

Table 2. Calculation of average Moodle storage by year and student

Megabyte used per year and per student	Gigabyte consumed by students since 2010	Terabyte consumed since 2010
89.3 MB	0.62 GB	5 TB

3.2 Characteristics of Big Data, Smart Data and the Data Lake

Table 3 shows a comparison of the characteristics of big data, smart data and the data lake. It has taken as a reference the capacity of these platforms for data management where big data handles a large volume of data [14]. The Smart data, with respect to capacity, acts as a data filter taking only the most important ones and, on those, applies the techniques of data analysis. The data lake, with its capacity, shares the characteristics of a data warehouse with the advantage that it does not need a process of extraction, transformation and load (ETL) that is in charge of a pre-processing of the data [18].

Table 3. Analyses of datasets

Platform	Capacity	Sources	Storage cost
Big data	Large dataset	Complex, structured or unstructured data set	High
Smart data	It is a data filter	Raw data	Medium
Data lake	Improved version of the data warehouse	Raw data	Low

With regard to the sources each of these supports, all three have similar characteristics, they feed on structured or unstructured datasets. Another feature that is important is the cost of storage where the big data has the highest cost due to the large volume of data and the infrastructure it supports for data acquisition. On the other hand, smart data, acting as a data filter, manages the storage cost better. Finally, a data lake is based on technologies that allow the storage of raw data and then apply incrementally the structure, as defined by the analytical requirements.

So far, we have detailed the characteristics of each of the data platforms which are important for the current situation of our platform, Moodle. The storage of the LMS of the university that has been used so far, since 2010, is 5 TB. The concept of big data is not considered as a fixed parameter from which it can be adopted as a big data platform. However, we will adopt as a reference a starting point of 30 TB [14]. If we consider the level of processing, speed and analysis, it is an advantage to use big data. However smart data and the data lake that have been considered in this work can also offer these characteristics without oversizing the resources. Therefore, categorisation with respect to storage is sufficient to rule out the adoption of big data for the needs of the institution.

The need is to convert the data into useful information. Considering this objective, we can better analyse both the adoption of smart data and a data lake. We begin this analysis by indicating that, in the case study of the Moodle platform, data mining techniques have been applied with the help of a data warehouse and the generation of cubes in the past. However, the life cycle of this technique has come to an end. With this consideration, at present the University has an original Moodle repository (MySQL) and a repository with clean data in an SQL database engine. In addition to these data sources should be considered external sources that are spreadsheets, plain text and even independent databases. These databases may contain relevant information from several of the courses and have been handled internally by teachers either as learning activities or records. As an additional point, the big data in our case is oversized by the volume of data available on the LMS platform, but we can benefit from the tools that big data offers for the analysis of the information.

3.3 Processing in Smart Data and a Data Lake

The smart data is not focused on storing or processing information, but on extracting value from it. This is where the human factor, the business knowledge and the expertise are most relevant. This task is achievable by data scientists who know the data they have, what they need to know and how to obtain it [13]. The questions to be answered in the analysis are: what do we do with all this volume?, what is the relevant

information of all that we have collected?, what level of aggregation is needed?. With regard to speed, we must know precisely what actions make sense in real time, which in near-real time and which can be performed every hour, every day, every month or every year. It does not make sense to analyse the information every second if we can only act every twenty-four hours, we should simply store it to analyse after.

In data lake the access to the original information is direct and reduces the intermediate steps for its processing [12]. Sometimes, when a record is deleted, it may not be needed immediately, but after a while. Data that may not be useful today may be needed after a few months or even years. A data lake marks the differences as a non-pre-processed data storage system. However, it is necessary to consider a higher cost, both of technical means and of professional profiles that are able to manage it.

Table 4 describes the parameters evaluated in the smart data and data lake platforms. The storage parameter refers to the capacity of the platform to provide the user with the storage of the data. The smart data does not act as a storage platform because it is focused on extracting useful information. Their accuracy depends on the analysis of the data scientists in determining the exact times for the execution of processes. The data lake, on its own, acts as a data repository since its processing uses raw data, its operation is based on the conservation of the data. For cost parameters at the technical level, it is considered that the smart data does not need a great infrastructure since it uses the operational data sources to extract the information; by contrast, the data lake, requires greater infrastructure for data storage as well as systems for processing. In human cost, the two platforms require highly trained analysts and data scientists with extensive knowledge of the business and the data generated.

Table 4. Technical analysis of smart data and the data lake

Platform	Storage	Prosecution	Technical costs	Human cost
Smart data	Low	High availability	Medium	High
Data lake	High	High availability	High	High

4 Analysis of Results

The analysis made in Sect. 3 gives us a broader picture of the tool that we can use considering the needs that are presented by the educational institution used as a case study. It is worth mentioning that any change considered for improvement in education should be attached to the economic reality of the institution. With this clarification, the tool that is sought must meet the technical and quality requirements, as well as the technical and human costs.

In the first analysis carried out in Sect. 3.2, big data has been discarded: although its functionality is broad and applicable to each organization often it is not an optimal solution, since it is possible to oversize the utility of the organizations technical resources. Excessive resources allocated to the system will affect implementation costs; however, we can make use of the analysis tools for our process.

In reviewing the needs as explained in Sect. 3.1, we note that the LMS used manages a data warehouse. This data store can be reused as an external source and be

managed by both a smart data as a data lake. The results indicate that the data lake provides advantages to the educational environment if it meets several characteristics, such as high availability in storage resources. In return, it offers us the conservation of data which, if not needed at this moment, but it may be important in the future. With this option, the analysis of several sensors or systems that help the discovery of trends or patterns of the students can be integrated. For example, whether it is necessary to increase the consumption of coffee in the period of examinations, or how many times a student enters the university, the data collected from his access card. The qualities offered by a data lake are interesting and give long-term advantages. However, at the moment our study only covers the LMS platform, so our application focuses on the use of smart data by reducing costs.

The use of smart data focuses on how we can integrate our data warehouse into its processing. The ideal for this tool is that it can make use of the cubes that are available in the current system without the need to process the information. It will simply extract the value of it into the process. Keeping data that has gone through a previous processing ensures the accuracy of data in the same way as it reduces processing. Another feature of smart data is the configuration of processes at specific times. For example, every 24 h, once a month or every 4 months, depending on the need of the organization.

The benefits offered by data mining for the analysis of data will be used, because this converges, without any problem, with the data and information generated by smart data. In this work, we do not perform an analysis of the algorithms to be used. However, from experience of the authors in previous works [15, 16], it can be mentioned that both the search algorithm and the cluster have sufficient characteristics to solve our needs.

On average one student generates 90 Mb per year, this volume of information is only from the use of the Moodle platform. The amount of information is very low in consideration of common enterprises, but this figure will increase if more sensors and systems are integrated into our analysis platform. The description of the problem mentions the need for integration of the systems that the university manages in order to carry out an in-depth analysis of the students. These systems manage the student's attendance, financial situation, qualifications, even there are printing systems that will indicate the trend in each student's reading.

5 Conclusions

This work includes new concepts that can be considered as a component that helps to improve education through the use of information and communication technologies. What has been sought during this development is to qualify the various platforms based on the needs of a particular educational institution considering, as the main base, the multiple sources of data.

Most organizations currently seek to take advantage of the information that is generated daily in the interaction with customers, defining what their interests are and being able to generate more profits based on these statistics. The same concept can be replicated in the educational field and, in this way, a personalised education can be offered based on the characteristics or patterns presented by each individual student.

The use of data mining on educational platforms every day has greater depth. However, it is important that the evaluation environment goes beyond an LMS. It is important that all the sensors or systems surrounding the student converge into one, so that the trends, problems or help that each student requires may be detected in a timely manner. This processing capacity exceeds the typical data warehouse so we have considered it important to scale to other types of tools.

Using a smart data will have the ability to analyse these systems in depth and establish patterns that tell us how the learning outcomes of a specific student can be improved. At the moment, a test on the operation of a smart data has been carried out using Microsoft power BI tool. The results will be presented in a future work since the data obtained are in validation stage. The power BI tool allows an ad hoc analysis and until now, four systems that control the student's activity when he or she is at university have been integrated into the test.

References

1. Dalsgaard, Ch.: Social software: e-learning beyond learning management systems. Eur. J. Open Distance E-Learn. **9**(2), 1–7 (2006)
2. Davenport, T.H., Barth, P., Bean, R.: How big data is different. MIT Sloan Manage. Rev. **54** (1), 4346 (2012). https://search.proquest.com/docview/1124397830?accountid=33194
3. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Mag. **17**(3), 37 (1996)
4. Higdon, S.J., Devost, D., Higdon, J., Brandl, B., Houck, J., Hall, P., Green, J.: The SMART data analysis package for the infrared spectrograph* on the spitzer space telescope. Publ. Astron. Soc. Pac. **116**(824), 975 (2004)
5. Lavalle, S., Lesser, E., Shocley, R.: Big data, analytics and the path from insights to value. MIT Sloan Manage. Rev. **52**(2), 21 (2011)
6. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.: Big data: the next frontier for innovation, competition, and productivity, pp. 27–36 (2011). http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation
7. Dougiamas, M., Taylor, P.: Moodle: using learning communities to create an open source course management system. In: Proceedings of ED-MEDIA World Conference on Educational Multimedia, Hypermedia and Telecommunications, pp. 171–178. Association for the Advancement of Computing in Education, Honolulu (2003)
8. O'leary, D.: Embedding AI and crowdsourcing in the big data lake. IEEE Intell. Syst. **29**(5), 70–73 (2014)
9. Sagiroglu, S., Sinanc, D.: Big data: a review. In: International Conference on Collaboration Technologies and Systems (CTS), pp. 42–47 (2013)
10. Snijders, C., Matzat, U., Reips, U.: Big Data: big gaps of knowledge in the field of internet science. Int. J. Internet Sci. **7**(1), 1–5 (2012)
11. Terrizzano, I.G., Schwarz, P.M., Roth, M., Colino, J.E.: Data wrangling: the challenging Yourney from the wild to the lake. In: Conference on Innovative Data Systems Research (CIDR), pp. 1–9 (2015)

12. Thusoo, A., Shao, Z., Anthony, S., Borthakur, D., Jain, N., Sen Sarma, J., Liu, H.: Data warehousing and analytics infrastructure at Facebook. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, pp. 1013–1020. ACM (2010)
13. Trautsch, F., Herbold, S., Makedonski, P., Grabowski. J.: Adressing problems with external validity of repository mining studies through a smart data platform. In: Proceedings of the 13th International Conference on Mining Software Repositories MSR, pp. 97–108. ACM (2016)
14. Villars, R.L., Carl, W., Matthew, E.: Big data: what it is and why you should care. White Paper IDC **14**, 1–14 (2011)
15. Villegas-Ch, W., Luján-Mora, S.: Systematic review of evidence on data mining applied to LMS platforms for improving e-learning. In: International Technology, Education and Development Conference (INTED), pp. 6537–6545 (2017)
16. Villegas-Ch, W., Luján-Mora, S.: Analysis of data mining techniques applied to LMS for personalized education. In: World Engineering Education Conference (EDUNINE), pp. 85–89. IEEE (2017)
17. Walker, J.S.: Big data: a revolution that will transform how we live, work, and think. Int. J. Advertising **33**(1), 181–183 (2014)
18. Widom, J.: Research problems in data warehousing. In: Proceedings of the Fourth International Conference on Information and Knowledge Management (CIKM), pp. 25–30 (1995)