ISSN: 1646-9895

# risti

Revista Ibérica de Sistemas e Tecnologias de Informação
Iberian Journal of Information Systems and Technologies

Abril 19 • April 19

Nº E19

Revista Ibérica de Sistemas e Tecnologias de Informação
Iberian Journal of Information Systems and Technologies

**Edição / Edition**
Nº. E19, 04/2019

**ISSN:** 1646-9895

**Indexação / Indexing**
Academic Journals Database, CiteFactor, Dialnet, DOAJ, DOI, EBSCO, GALE, Index-Copernicus, Index of Information Systems Journals, Latindex, ProQuest, QUALIS, SCImago, SCOPUS, SIS, Ulrich's.

**Propriedade e Publicação / Ownership and Publication**
AISTI – Associação Ibérica de Sistemas e Tecnologias de Informação
Rua Quinta do Roseiral 76, 4435-209 Rio Tinto, Portugal
**E-mail:** aistic@gmail.com
**Web:** http://www.aisti.eu

Luis David Narváez. Pontificia Universidad Católica del Ecuador Sede Ibarra (PUCESI). Ecuador.

María Alejandra Noguera. WSP. Chile.

Paola Ordoñez. Pontificia Universidad Católica del Ecuador Sede Ibarra (PUCESI). Ecuador.

Yadira Ordoñez. Pontificia Universidad Católica del Ecuador Sede Ibarra (PUCESI). Ecuador.

Miguel Ángel Orosa. Pontificia Universidad Católica del Ecuador Sede Ibarra (PUCESI). Ecuador.

Gustavo Paredes. Universidad de Los Andes (ULA). Venezuela.

Susana Patiño. Pontificia Universidad Católica del Ecuador Sede Esmeraldas (PUCESE). Ecuador.

Anna Pérez. Universidad de Los Andes (ULA). Venezuela.

Rubén Pérez. Gas Energy. Venezuela.

Pablo Pico. Pontificia Universidad Católica del Ecuador Sede Esmeraldas (PUCESE). Ecuador.

Ernesto Ponsot. Universidad de Los Andes (ULA). Venezuela.

Iván Puentes. Universidade de Vigo. España.

Marco Pusdá. Universidad Técnica del Norte (UTN). Ecuador.

José Antonio Quiña Mera. Universidad Técnica del Norte (UTN). Ecuador.

Xavier Quiñonez. Pontificia Universidad Católica del Ecuador Sede Esmeraldas (PUCESE). Ecuador.

Yalitza Ramos Gil. Pontificia Universidad Católica del Ecuador Sede Ibarra (PUCESI). Ecuador.

Edmundo Recalde. Pontificia Universidad Católica del Ecuador Sede Ibarra (PUCESI). Ecuador.

Carlos Rivas. Unidad Doc. de Medicina Familiar y Comunitaria. España.

Francklin Rivas. Universidad Técnica Federico Santa María (UTFSM). Chile.

Néstor Diego Rivera Campoverde. Universidad Politécnica Salesiana. Ecuador.

Dulce Rivero. Pontificia Universidad Católica del Ecuador Sede Ibarra (PUCESI). Ecuador.

Francisco Rodríguez. Universidad de Los Andes (ULA). Venezuela.

Magdalena Rodríguez Fernández. Universidad de la Coruña. España.

Ana Isabel Rodríguez Vázquez. Universidade de Santiago de Compostela (USC). España.

Revista Ibérica de Sistemas e Tecnologias de Informação
Iberian Journal of Information Systems and Technologies

Recebido/Submission: 16/12/2018
Aceitação/Acceptance: 03/02/2019

# Management of educative data in university students with the use of big data techniques

William Villegas-Ch.[1], Xavier Palacios-Pacheco[2], Iván Ortiz-Garcés[3], Sergio Luján-Mora[4]

**william.villegas@udla.edu.ec, xpalacio@uide.edu.ec, ivan.ortiz@udla.edu.ec, sergio.lujan@ua.es**

[1,3] Universidad de Las Américas, 170523, Quito, Ecuador.

[2] Universidad Internacional del Ecuador, 170411, Quito, Ecuador.

[4] Universidad de Alicante, 03080, Alicante, España.

**Pages:227–238**

***Abstract:*** The large volumes of data that exist in the universities keep important information of each student. Analyzing this data represents a challenge for data scientists due to the number of resources they consume. Many of the universities do not have the capacity of infrastructure as well as human resources to do it for this reason they desist from the analysis of data depriving themselves of generating knowledge about their students. The range of sensors that generate data in a university is so wide that doing an analysis of data through a traditional method such as business intelligence does not provide accurate results and their response times are not as expected. This work proposes the use of big data techniques in a university to obtain accurate results in real time that will help in making decisions improving education and learning.

***Keywords:*** big data; Hadoop; analysis of data.

## 1. Introduction

Universities, like companies with their clients, make decisions for their students based on the data they have about them. However, this process increasingly requires methods that allow an analysis of the data superior and that the results appear at appropriate times. This does not mean that the universities have not already been doing work on the data and getting knowledge. The problem lies in the large volume of data generated from student activities that greatly exceed the capabilities of classical analysis platforms such as business intelligence (BI). The variety of sources and that these do not respond to a strictly structured data model leads to the search for other alternatives, as well as concepts in data analysis (Cheng & Cheng, 2011).

One of these alternatives and that is marked, as a trend in the analysis is the use of big data. These platforms offer alternatives for the treatment of data and that obtaining knowledge is more flexible, with lower costs and in shorter times. For example, in a common and very important analysis for universities is to detect and classify students who have learning problems and which leads to high dropout rates. With the use of business intelligence and data mining it can be done, however, many data that are not structured are left aside (Villegas-Ch, Lujan-Mora & Buenano-Fernandez, 2018). The experience in the management of these tools suggests that the greater the sources considered in the analysis closest to reality are the results. These unstructured sources usually come from the students' navigation log, their Internet searches, and their interaction in social networks. All this data is impossible to leave out since the use of information and communication technologies (ICT) are an active part of the students (Li, Zhang & Wang, 2013). Being able to draw trends or determine the best way to learn from a student benefits all the components within an educational environment.

This paper considers the use of big data as an alternative to data analysis of a university where there are great diversity sources. The purpose is to consider the steps that an institution should consider to include in its processes the integration of these tools, as well as the use of the Hadoop framework as a manager in the analysis of data. The work is divided as follows: section 2 presents the concepts used for the development of the method; section 3 contains the method where the different phases to be considered for the implementation of a big data platform is established; section 4 presents the conclusions found in the development of the work.

## 2. Preliminary concepts

### 2.1. Big data

Big data is the massive data analysis, an amount of data, so large, that traditional data processing software applications are not able to capture, process and present results in a reasonable time (Shah, Soriano, & Coutroubis, 2017). Big Data was born with the aim of covering needs not met by existing technologies. In education, it can have an important impact on teachers, school systems, students and curricula. The analysis of big data can identify students at risk, ensure that students have adequate progress and can implement a better system for the evaluation and support of teachers and principals (Villegas-Ch et al., 2018). To comply with this process, big data techniques work in the storage and processing of large volumes of data that have specific characteristics such as:

- Volume refers to the size of the data that can come from multiple sources.
- Speed defines the speed with which the data arrive using units such as tera, peta or exabytes.
- Variety, we speak of data, structured, Semi-structured, Unstructured.

### 2.2. Data sources

The information available in universities has grown exponentially due to the inclusion of ICT in education (Cong & Xiaoyi, 2009). The data of this interaction is stored in multiple data sources that support academic management. Next, the following stand out:

- Produced by people. Send an email, write a comment on Facebook, answer a survey, enter information in a spreadsheet, use the learning management systems (LMS) and click on an Internet link. These actions, which are basic and carried out on a daily basis, represent an immense source of data.

- Between machines. The machines share data directly; this action is known as machine-to-machine (M2M). Thus, the parking meters mobile phones, vending machines for drinks and food in the university, to put a few examples, communicate through devices with other machines (Datta & Bonnet, 2014). The transmitted data is stored in different repositories. The communication networks to carry out these actions are very varied. Among the best known are Wifi, ADSL, fiber optics and Bluetooth.
- Biometrics. The data may originate from fingerprint sensors, retinal scanners, DNA readers, face recognition sensors or speech recognition. Its use is common in terms of safety in all its variants.
- Web marketing. All movement in the network is subject to all types of measurements that have marketing studies and behavioral analysis. With the analysis of these data, one can conclude the trends of each student (Bengel, Shawki, & Aggarwal, 2015). For example, the websites most accessed by students or places where they spend more time.

### 2.3. Type of data

In addition to their origin, the data can have classified into three classes according to their structure: Unstructured data types: documents, videos, audios, etc.; semi-structured data types: software, spreadsheets, reports; types of structured data. Only a small percentage of the information is structured and that can cause many errors if a data quality process is not applied (Gubanov, Stonebraker & Bruckner, 2014).

## 3. Method

The university that participates in this study stores a large amount of data. The majority of this data comes from the activities carried out by students in the LMS, academic management systems and sensors located in the university facilities. The sensors acquire and store all kinds of information about the activities that students perform within the university. This information in conjunction with all systems that are responsible for academic management can promote an important redesign in the learning methods used today.

When a big data process is included for the first time in a university it is important to generate working groups that prepare all the actors for the future change. The characteristic of this change is that the main value of big data does not come from the data in its raw form, but from its processing and analysis and from the insights, products, and services emanating from the analysis (Laigner et al., 2018). Radical changes in technologies and management methods are similarly dramatic changes in the way data supports decisions and innovation in products/services.

### 3.1. Phases of big data

The analysis prior to the implementation of a big data model is to determine the budget allocated during the process and the resources that will intervene considering the following parameters (Mohammed, Humbe & Chowhan, 2016):

- Managers are the sponsors, project managers, coordinators and quality managers immersed in an educational environment.
- Designers and data architects have technical profiles with clear objectives regarding the implementation of the project.
- Implementers, is the qualified personnel, analysts and developers, with knowledge of the sector and technology.
- Data operators are the analysts in charge of data at the entry, intermediate and result level.

Once the budget is determined, it passed to the design phase according to the needs of the university and it optimized considering the cost, the scalability and the different options of the market. It consists of two stages:

- Infrastructure, are networks, computers or servers, that is, the physical support of the solution.
- Architecture is the logical support of the solution, formed by protocols, communications or procedures.

For the implementation of the big data model in addition to the phases indicated, it is important to define aspects such as administration, maintenance, and security. The steps, for this, are:

- Installation of servers, components and start-up of the infrastructure.
- Configuration of the infrastructure for its correct functioning.

Figure 1 shows the phases that allow the execution of big data where two processes are considered. The first process is the engineering of big data that is composed of the acquisition and preparation of data (Chen, Mao, & Liu, 2014). The second process is the big data analytics, which consists of analyzing the data, reporting and acting.

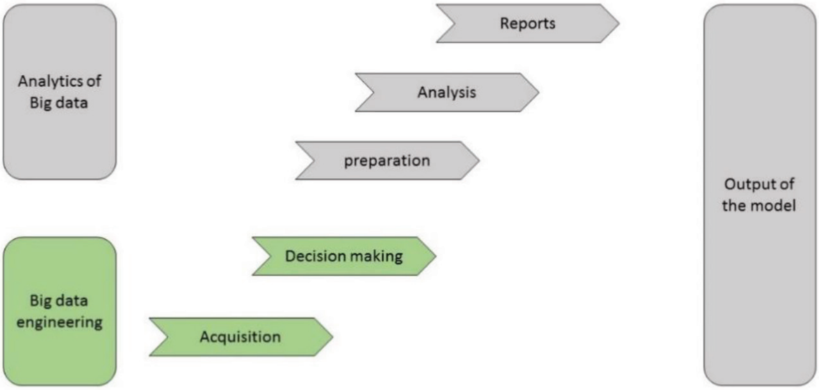

Figure 1 – Phases of execution of big data

### 3.2. Acquisition (big data engineering)

The first step in the application of big data in a university environment is to understand where the data comes from (Arruda & Madhavji, 2017). For which, the selection of sources considers three categories that fit this type of environment:

- Streaming data. It includes data that reaches the ICT systems of a network of connected devices. The data is analyzed as they arrive, this process determines which data to keep and which requires a deeper analysis.
- Social media data. Social interaction data is an attractive set of information, particularly for marketing, academic monitoring, and support functions. These data often have unstructured or semi-structured formats, so they present a unique challenge when it comes to consumption and analysis.
- Fonts machine to machine. The data of the different sensors contribute to the process identifying tendencies, places and times of permanence of the students in the university.

### 3.3. Preparation (big data engineering)

For the preparation of the data, certain guidelines are considered that make it possible to take full advantage of the information. The clarity in the guidelines supports the preparation of the data, the amount of data considered and how to use the knowledge discovered.

Raw data extracted directly from the sources are never in the format needed to analyze them. To solve the problem, it is important to prepare the data by applying two main objectives (Taleb & Serhani, 2017). The first is to purge the data to address data quality problems, and the second to transform the raw data to adapt it to the analysis.

The refinement identifies the erroneous data, corrects them or eliminates them; this process allows to improve the data and to consider only those that present a certain level of quality. There are quality problems with data from applications that are in production that include inconsistent data, duplicate records, missing data, etc. For example, the address of a student registered in two different places. This is an example of records that do not match; the lack of demographic data, unavailable values such as lack of a student’s age, invalid data. For example, a telephone number, and outliers that cause values to be much higher or lower than expected. The quality problems in the data are solved by detection and correction of errors. In the correction, there are several methods; all depend on a previous analysis that allows applying to correct them. One of these methods is the elimination of records with unavailable values. Another method is to combine the duplicate records presenting unique and validated information.

The correction of the invalid values is made by replacing these values with the best estimate of a fair value. For example, for a missing value in the field of a student’s age, the semester the student attends is estimated as a fair value. Outliers can have deleted as long as they are not important for the analysis. An effective process in the correction of errors implies having the clear knowledge of the application (Sehgal & Agarwal, 2016). For example, how the data has been collected, the population and the intended use of the application. This knowledge of the domain is essential to make decisions about how to handle incomplete or incorrect data.

The second part of the data preparation is to manipulate the purified data to convert it to the format needed for the analysis (Londhe & Rao, 2017). Some operations in this phase include scaling, transformation, feature selection, dimensionality reduction, and data manipulation.

- Scaling involves changing the range of values so that it is within a specific range, such as from zero to one. This is done to prevent certain features with large values from dominating the results.
- The transformation in the data eliminates noise and variability. The result of the transformations are considered as aggregate data. Adding data to the process results in a decrease in variability, which helps the analysis.
- The selection of data involves the elimination of redundant or irrelevant features, the combination or creation of new features. During the data exploration phase, it is possible to discover that some characteristics are correlated. In this case, it is possible to eliminate one of the characteristics without adversely affecting the results of the analysis.
- The reduction of dimensionality is useful when the data set has a large number of dimensions. The reduction allows finding a smaller subset of dimensions that captures most of the variation in the data.
- The manipulation is a way to verify that the raw data is in the correct format for the analysis. For example, from samples that record semi-annual changes in student grades, it is possible to capture changes in student performance for a given cycle. With this information, they can be grouped and calculate the mean, range and standard deviation for each group.

The preparation stage includes the socialization of the solution to the different departments responsible for the educational management and obtains the relevant authorizations to be able to carry it out. This stage includes:

- Detect the needs, have to do with the volume of data to be stored, its variety, the speed of collection, processing, and horizontal scalability. This process also reveals shortcomings when confronting the new technology with the existing one in the university.
- Justify the investment, big data improves the identification of both academic and technical problems and this by creating a high-performance environment that enables cost savings and improvement in academic quality.
- Evaluate the limitations; consider the infrastructure, technological maturity, resources, even the legal aspects in relation to data privacy.

### 3.4. Analysis (analytics of big data)

Data analysis involves constructing a model from the input data using an analysis technique to generate the output data (Eluri et al., 2016). There are different types of problems and different types of analysis techniques such as classification, regression, clustering, association analysis, and graphical analysis (Agnihotri & Sharma, 2015).

- The classification predicts the category of the input data. For example, predict the possible problems that a student encounters in the development of mathematical exercises.
- The regression predicts a numerical value instead of a category. For example, the prediction of the qualification of a questionnaire. The rating is a numerical value, not a category, so it is a regression task instead of a classification task.

- Clustering is organizing similar elements in groups. For example, group the base of students in different segments to recommend activities in such a way that learning is more effective.
- The association consists of developing a set of rules to capture associations within elements or events. Rules are used to determine when elements or events occur at the same time. For example, the association analysis may reveal that students who have good grades also tend to be interested in extracurricular activities.
- Graphical analysis to analyze data occurs when there is a large number of entities and connections between these, as in social networks. For example, in the graphic analysis, it may be useful to study the performance of a student over a period.

### 3.5. Reports (analytics of big data)

The potential of Big Data lies in the analysis and in converting the data into relevant information. Here comes into play the use of different techniques such as data mining and methodologies based on machine learning. These techniques and methodologies are those that can extract the true value to the information (Cao & Gao, 2018). The reports and the way in which the information is presented manage to complete the transformation of the commercial model created in previous phases into one or more representations of specific data of the university. Once the reports are obtained and with the complete modeling, it is necessary to evaluate and validate the results. The evaluation of the results depends on the type of analysis techniques used. For example, to get an idea of the model's performance with the data, it is evaluated based on questions such as:

- Is it necessary to perform the analysis with more data in order to obtain a better performance of the model?
- Does it help to use different types of data?
- Is it difficult to distinguish students with different needs in the results of clustering?
- Does adding the zip code to the input data help generate more granular student segments?
- Do the results of the analysis suggest a more detailed view of some aspects of the problem?

For example, the prediction of the percentage of students who pass mathematics gives good results, but the predictions of the results in the subject of calculation are not good. In the second case, the samples of the academic activities in the matter of calculation need deeper analysis. The factors that influence the results can be as diverse as for example; the existence of anomalies in the sample or the need to include additional data to fully capture the students' performance. What is sought is that the model is effective with respect to the success criteria defined at the beginning of the project. In this case, communication and action must be prepared on the results obtained in the analysis.

### 3.6. Decision making (analytics of big data)

The decision making works in tandem with the previous phase since after the analysis the conclusions come to carry out actions and make decisions. The final goal of data analysis is to execute new strategies that improve academic management and student learning. The premise is that the analysis is in real time and as quickly as possible. The results obtained in the analysis convert the raw data into “actionable knowledge”. For an adequate and understandable management of the different areas, it is necessary to integrate visualization tools that conceive the reality of the study environment or be able to predict the future.

### 3.7. Data management with Hadoop

For data management, considering the different existing sources it is important to use tools capable of carrying out the process at the appropriate times. For this work, Hadoop is used as an open source framework to store data and run applications in clusters. Provides massive storage for any type of data, enormous processing power and the ability to process virtually unlimited concurrent tasks or works (Mazumdar & Dhar, 2015).

The architecture of Hadoop allows carrying out an effective analysis of large volumes of data, adding a value helps to make strategic decisions, to improve educational processes. This architecture allows to monitor what the students think or to draw scientific conclusions about the learning problems presented by different groups of students. With Hadoop, universities can explore complex data through customized analysis tailored to their students and needs.

The architecture of Hadoop is composed of three fundamental pillars that make it a versatile tool, flexible and fault tolerant. Its pillars are a distributed file system, called HDFS for its operation. The Hadoop engine consists of a MapReduce job scheduler, as well as a series of nodes responsible for executing them (Bhandarkar, 2010). A set of utilities that make possible the integration of subprojects.

On the file system is the MapReduce engine, which is a job scheduler, called JobTracker that is responsible for sending jobs to the nodes. MapReduce sends the incoming workflow to the TaskTracker nodes available in the cluster that are responsible for executing the map functions and reduces in each node. The planner keeps those jobs as close to the machine that has issued that information as possible. If the work cannot be located in the current node in which the information resides, the nodes in the same rack are given priority. This allocation reduces network traffic on the cluster’s core network. If a TaskTracker fails or suffers a waiting time, that part of the work is reprogrammed. Hadoop responds to a master-slave structure where the JobTracker is located in the master while there is a TaskTracker for each slave machine as shown in Figure 2. The JobTracker records the pending jobs that reside in the file system. When a JobTracker starts, it looks for that information, so that it can start the work again from the point where it was left.

The Hadoop file system (HDFS) handles two fundamental elements in the architecture: The NameNode and the DataNode. The NameNode is only found in the master node and is responsible for keeping all the stored data indexed. That is, the application needs specific information about the location of the data. The NameNode is found in the slaves and is responsible for storing the data.

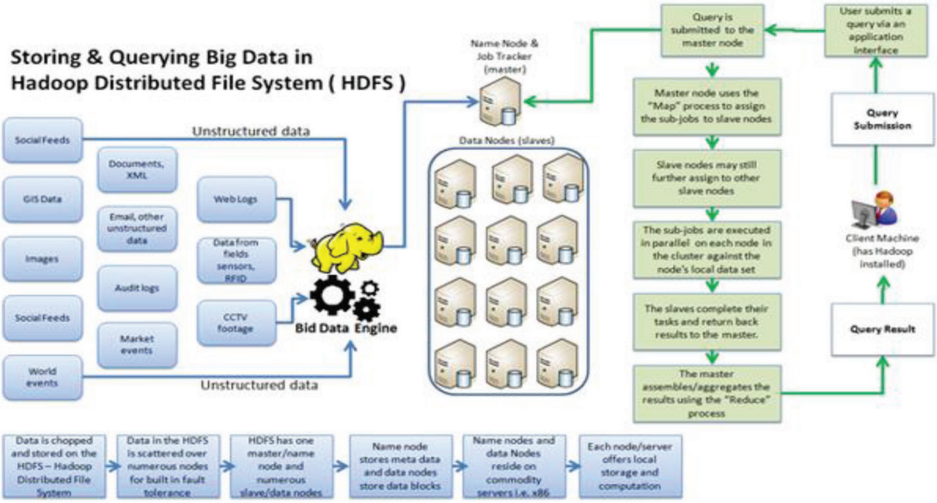
Storing & Querying Big Data in
Hadoop Distributed File System ( HDFS )
Social Feeds
GIS Data
Images
Social Feeds
World events
Documents, XML
Email, other unstructured data
Audit logs
Market events
Web Logs
Data from fields sensors, RFID
CCTV footage
Unstructured data
Unstructured data
Bid Data Engine
Name Node & Job Tracker (master)
Data Nodes (slaves)
Query is submitted to the master node
User submits a query via an application interface
Master node uses the "Map" process to assign the sub-jobs to slave nodes
Slave nodes may still further assign to other slave nodes
The sub-jobs are executed in parallel on each node in the cluster against the node's local data set
The slaves complete their tasks and return back results to the master.
The master assembles/aggregates the results using the "Reduce" process
Query Submission
Client Machine (has Hadoop installed)
Query Result
Data is chopped and stored on the HDFS – Hadoop Distributed File System
Data in the HDFS is scattered over numerous nodes for built in fault tolerance
HDFS has one master/name node and numerous slave/data nodes
Name node stores meta data and data nodes store data blocks
Name nodes and data Nodes reside on commodity servers i.e. x86
Each node/server offers local storage and computation

Figure 2 – Hadoop architecture (Ticout Outsourcing Center, 2013)

### 3.8.Hadoop MapReduce

Hadoop provides an execution environment oriented to applications developed under the MapReduce programming paradigm (Merla & Liang, 2017). Under this model, the execution of an application presents two stages:

- Map: where ingestion and transformation of the input data takes place, in which the input registers are processed in parallel.
- Reduce: aggregation or summary phase, where all the associated records are processed by the same entity.

The main idea, on which the Hadoop MapReduce execution environment revolves is that the entry is divided into fragments and, each fragment, is treated independently by a map task. The results of processing each fragment are physically divided into different groups. Each group is sorted and goes to a task reduced.

The execution cycle of an application in Hadoop is shown schematically. The developer only provides four functions to the Hadoop framework: the function that reads the input records and transforms them into tuples (RecordReader), the map (Mapper) function, the reduce (Reducer) function, and the function that transforms the pairs generated by the function reduces in output registers (RecordWriter).

### 3.9.Application cycle

The execution of an application in Hadoop consists of the presentation of the application, the generation of the ApplicationMaster for the application and the execution of the application managed by the ApplicationMaster. For example, a client program sends the request, including the necessary specifications to launch the application ApplicationMaster itself. The ResourceManager assumes the responsibility of negotiating a container in which the ApplicationMaster must start, and then executes the ApplicationMaster. The ApplicationMaster, once started, is registered with the ResourceManager.

The registry allows the program to consult details of the resources to the ResourceManager. During normal operation of the execution, the ApplicationMaster negotiates containers of appropriate resources through the resource protocol. When the container assignment is satisfactory, the ApplicationMaster launches the container, providing the specifications of the container execution to the NodeManager. The execution information, in general, includes the necessary information to allow the container to communicate with the same ApplicationMaster.

The code of the application that runs inside the container provides the necessary information (progress, status, etc.) to your ApplicationMaster through a specific protocol of the application. During the execution of the application, the client that presented the program communicates directly with the ApplicationMaster to check the status, progress updates, etc. Through a specific protocol of the application. Once the application is completed and all work is completed, ApplicationMaster cancels the registration with the ResourceManager and closes, allowing the container to be reused for another application (Verma & Pandey, 2016).

## 4. Conclusions

This work is a detail of the steps that must be followed for the application of a big data platform. This method is in the testing phase where a large amount of data from different types of sources has been added. The exercise before going to production requires the comparison of results and the evaluation of all phases. So far it has not been possible to perform the evaluation because Hadoop needs deep knowledge in Java to be able to correctly integrate structured data. However, it has been possible to carry out tests in the different phases and the speed of processing, as well as the high savings in infrastructure, are details that allow us to continue with the investigation.

Several methods and models help the implementation of big data in a university institution. The first and most important step in the process is the location, analysis and data cleansing, this has taken more than 60% of the project's execution time.

Having a tool that allows knowing the students' tendency and the way they learn in a university environment is necessary for decision making. It must be borne in mind that in order to carry out this analysis what data scientists are looking for is that there is no ambiguity in the data. For this reason, consider the data that students generate in their normal activities without these being the results of surveys or questionnaires where the answers can be biased, it is an advantage over traditional data analysis platforms.

## References

Agnihotri, N., & Sharma, A. K. (2015). Proposed algorithms for effective real time stream analysis in big data. In *2015 Third International Conference on Image Information Processing (ICIIP)* (pp. 348–352). IEEE. DOI: 10.1109/ICIIP.2015.7414793

Arruda, D., & Madhavji, N. H. (2017). Towards a requirements engineering artefact model in the context of big data software development projects: Research in progress. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 2314–2319). DOI: 10.1109/BigData.2017.8258185

Bengel, A., Shawki, A., & Aggarwal, D. (2015). Simplifying web analytics for digital marketing. In *2015 IEEE International Conference on Big Data (Big Data)* (pp. 1917–1918). IEEE. DOI: 10.1109/BigData.2015.7363968

Bhandarkar, M. (2010). MapReduce programming with apache Hadoop. In *2010 IEEE International Symposium on Parallel & Distributed Processing (IPDPS)* (p. 1). IEEE. DOI: 10.1109/IPDPS.2010.5470377

Cao, R., & Gao, J. (2018). Research on reliability evaluation of big data system. In *2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)* (pp. 261–265). IEEE. DOI: 10.1109/ICCCBDA.2018.8386523

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

Cheng, L., & Cheng, P. (2011). Integration: Knowledge Management and Business Intelligence. In *2011 Fourth International Conference on Business Intelligence and Financial Engineering (BIFE)* (pp. 307–310). IEEE. DOI: 10.1109/BIFE.2011.172

Cong, P., & Xiaoyi, Z. (2009). Research and Design of Interactive Data Transformation and Migration System for Heterogeneous Data Sources. In *2009 WASE International Conference on Information Engineering (ICIE)* (pp. 534–536). IEEE. DOI: 10.1109/ICIE.2009.222

Datta, S. K., & Bonnet, C. (2014). Smart M2M Gateway Based Architecture for M2M Device and Endpoint Management. In *2014 IEEE International Conference on Internet of Things(iThings), and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing(CPSCom)* (pp. 61–68). IEEE. DOI: 10.1109/iThings.2014.18

Eluri, V. R., Ramesh, M., Al-Jabri, A. S. M., & Jane, M. (2016). A comparative study of various clustering techniques on big data sets using Apache Mahout. In *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)* (pp. 1–4). IEEE. DOI: 10.1109/ICBDSC.2016.7460397

Gubanov, M., Stonebraker, M., & Bruckner, D. (2014). Text and structured data fusion in data tamer at scale. In *2014 IEEE 30th International Conference on Data Engineering (ICDE)* (pp. 1258–1261). IEEE. DOI: 10.1109/ICDE.2014.6816755

Laigner, R., Kalinowski, M., Lifschitz, S., Salvador Monteiro, R., & de Oliveira, D. (2018). A Systematic Mapping of Software Engineering Approaches to Develop Big Data Systems. In *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)* (pp. 446–453). IEEE. DOI: 10.1109/SEAA.2018.00079

Li, X., Zhang, F., & Wang, Y. (2013). Research on Big Data Architecture, Key Technologies and Its Measures. In *2013 IEEE International Conference on Dependable, Autonomic and Secure Computing (DASC)* (pp. 1–4). IEEE. DOI: 10.1109/DASC.2013.28

Londhe, A., & Rao, P. P. (2017). Platforms for big data analytics: Trend towards hybrid era. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)* (pp. 3235–3238). IEEE. DOI: 10.1109/ICECDS.2017.8390056

Mazumdar, S., & Dhar, S. (2015). Hadoop as Big Data Operating System -- The Emerging Approach for Managing Challenges of Enterprise Big Data Platform. In *2015 IEEE First International Conference on Big Data Computing Service and Applications (BigDataService)* (pp. 499–505). IEEE. DOI: 10.1109/BigDataService.2015.72

Merla, P., & Liang, Y. (2017). Data analysis using hadoop MapReduce environment. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 4783–4785). IEEE. DOI: 10.1109/BigData.2017.8258541

Mohammed, A. F., Humbe, V. T., & Chowhan, S. S. (2016). A review of big data environment and its related technologies. In *2016 International Conference on Information Communication and Embedded Systems (ICICES)* (pp. 1–5). IEEE. DOI: 10.1109/ICICES.2016.7518904

Sehgal, D., & Agarwal, A. K. (2016). Sentiment analysis of big data applications using Twitter Data with the help of HADOOP framework. In *2016 International Conference System Modeling & Advancement in Research Trends (SMART)* (pp. 251–255). IEEE. DOI: 10.1109/SYSMART.2016.7894530

Shah, S., Soriano, C. B., & Coutroubis, A. D. (2017). Is big data for everyone? the challenges of big data adoption in SMEs. In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 803–807). IEEE. DOI: 10.1109/IEEM.2017.8290002

Taleb, I., & Serhani, M. A. (2017). Big Data Pre-Processing: Closing the Data Quality Enforcement Loop. In *2017 IEEE International Congress on Big Data (BigData Congress)* (pp. 498–501). IEEE. DOI: 10.1109/BigDataCongress.2017.73

Ticout Outsourcing Center. (2013). Introducción a Hadoop y su ecosistema. *Ticout*. Retrieved from http://www.ticout.com/blog/2013/04/02/introduccion-a-hadoop-y-su-ecosistema/

Verma, C., & Pandey, R. (2016). Big Data representation for grade analysis through Hadoop framework. In *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)* (pp. 312–315). IEEE. DOI: 10.1109/CONFLUENCE.2016.7508134

Villegas-Ch., W., Lujan-Mora, S., & Buenano-Fernandez, D. (2018). Towards the Integration of Business Intelligence Tools Applied to Educational Data Mining. In *2018 IEEE World Engineering Education Conference (EDUNINE)* (pp. 1–5). IEEE. DOI: 10.1109/EDUNINE.2018.8450954

Villegas-Ch., W., Luján-Mora, S., Buenaño-Fernandez, D., & Palacios-Pacheco, X. (2018). *Big data, the next step in the evolution of educational data analysis. Advances in Intelligent Systems and Computing* (Vol. 721). DOI: 10.1007/978-3-319-73450-7_14