IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON MACHINE LEARNING DESIGNS, IMPLEMENTATIONS AND TECHNIQUES

Received January 1, 2020, accepted January 15, 2020, date of current version February 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2968969

Developing Big Data Projects in Open University Engineering Courses: Lessons Learned

JUAN A. LARA[1], (Member, IEEE), AUREA ANGUERA DE SOJO[2], SHADI ALJAWARNEH[3], (Member, IEEE), ROBERT P. SCHUMAKER[4], (Senior Member, IEEE), AND BASSAM AL-SHARGABI[5], (Member, IEEE)

[1]Escuela de Ciencias Técnicas e Ingenierías, Madrid Open University (UDIMA), 28400 Madrid, Spain
[2]Escuela Técnica Superior de Ingeniería de Sistemas Informáticos, Technical University of Madrid (UPM), 28031 Madrid, Spain
[3]Faculty of Computer and Information Technology, Jordan University of Science and Technology (JUST), Irbid 22110, Jordan
[4]Data Analytics Lab, University of Texas at Tyler (UT Tyler), Tyler, TX 75799, USA
[5]Faculty of Information Technology, Middle East University (MEU), Amman 11610, Jordan

Corresponding author: Juan A. Lara (juanalfonso.lara@udima.es)

ABSTRACT Big Data courses in which students are asked to carry out Big Data projects are becoming more frequent as a part of University Engineering curriculum. In these courses, instructors and students must face a series of special characteristics, difficulties and challenges that it is important to know about beforehand, so the lecturer can better plan the subject and manage the teaching methods in order to prevent students' academic dropout and low performance. The goal of this research is to approach this problem by sharing the lessons learned in the process of teaching e-learning courses where students are required to develop a Big Data project as a part of a final degree/master course. In order to do so, a survey was carried out among a group of students enrolled in those kinds of courses during the last years. The quantitative and qualitative analysis of the obtained data led us to present a series of lessons learned that may help other participants (both students and lecturers) to better study, design and teach similar courses. In addition, the results shed light on possible existing open problems in the area of Big Data project development. Both the methodology used and the survey designed in this research were validated by a group of experts in the area using a formal statistical approach at a significance level of $p<0.008$, which support the validity of the lessons learned.

INDEX TERMS Academic projects, big data, data mining, data science, NoSQL, lessons learned.

I. INTRODUCTION

The term *Big Data*, which was adopted in 2008 [1], refers to a discipline where large quantities of data are extracted, stored, visualized and analyzed in order to draw conclusions from them. The basis of this discipline was established several decades ago [2], but the term itself and its popularity has only appeared in the last few years [3]–[5]. The data to be analyzed is characterized by what was formerly called the three V's: Volume, Variety and Velocity [6] although nowadays, some authors also take other V's into consideration, such as Veracity [7].

There are multiple applications of big data in practically every field and discipline. Some examples are healthcare [8]–[10], Internet of Things [11]–[13], and the manufacturing industry [14]–[16].

The associate editor coordinating the review of this manuscript and approving it for publication was Oguz Bayat.

Apart from these applications, another field where Big Data can be of great use is in Education. The field of Education generates large amounts of data which must be analyzed with Big Data techniques in order to extract information [17], [18]. Additionally, there are an increasing number of courses that incorporate Big Data subjects as part of the students' curriculum [19], [20]. These tasks are especially interesting in the context of the research undertaken in this article, as the goal is to demonstrate how students can successfully undertake Big Data projects for their end-degree projects.

In Big Data projects and especially in those that are developed as part of a university assignment, ensuring quality is essential given that these projects must be presented and defended before a panel of experts on the topic. To succeed in completing such Big Data academic projects achieving a certain level of quality is not an easy task sometimes, since many difficulties and risks of many types (personal, educational,

technological, etc.) may arise. Based on our experience and according to the literature [21], those particular issues, if not managed properly, may lead to the student's dropout in a quite frequent rate.

To face this problem, this paper especially seeks to carry out an in-depth study of the nature, difficulties and challenges faced by students when undertaking Big Data projects as part of their curriculum. Its main contribution is to provide a series of lessons and recommendations that may help the different stakeholders involved in successfully implementing future Big Data project courses. It is especially aimed at:

a) Other students and professionals, as prior knowledge of the difficulties and characteristics of this type of projects will give them a greater chance of success and the chance to develop a better plan.
b) Lecturers who teach subjects where Big Data projects must be developed, as the conclusions drawn from this research will help them to boost training action in the most critical aspects of Big Data project development.
c) Educational institutions, which can better design and plan the curriculum that include Big Data topics and the necessary resources.

The scope of this contribution is limited to the knowledge that can be extracted from a single academic experience that is successful but low-scale, within the environment of an Open University and with data obtained from eight Big Data projects carried out by students in the past few years. Despite that, it is the first study of this type developed by the educational scientific community to the best of our knowledge.

The research methodology is based on the assumption that Big Data projects consist of the following stages: Data Collection, NoSQL Storage, Data Preparation, Data Analysis, Data Visualization; although it is true that not all projects may include each and every stage, or there might be projects that take other stages into account.

In order to address the problem stated above, our study attempts to respond to the following Research Questions (RQ):

- RQ1: What percentage of effort in a Big Data project is required for each stage?
- RQ2: What are the main risks/difficulties that may threaten a project?
- RQ3: What is the degree of difficulty of each stage in relation to the other stages?
- RQ4: What are the most frequently used technologies in each stage?
- RQ5: What is the stage that has the largest amount of available resources, tools, literature, etc. and what is the stage that has the least?
- RQ6: What aspects (recommendations) are key to achieving a successful Big Data project?
- RQ7: What existing problems/challenges must be faced by the Big Data community?

Note that these research questions may be divided into five major blocks: time required (RQ1), difficulties encountered (RQ2 and RQ3), resources (RQ4 and RQ5), recommendations drawn (RQ6) and challenges (RQ7). Once we know the answer to those RQ, we could extract some interesting lessons learned and understanding about this discipline that might be useful to mitigate the negative effects of big data project development complexity in students' academic success and performance. We have demonstrated the validity of our approach be means of a formal statistical approach based on the indications of a panel of experts in the area of academic big data project development.

The rest of this document is organized as follows: Section II focuses on the research methodology. Section III provides a description of the Big Data projects on which this research is based. Section IV lists the obtained results and the discussion. Section V present the validation carried out with experts. And finally Section VI presents the conclusions of this study and proposes future lines of research.

II. METHODOLOGY

This research is based on a total of eight projects undertaken by final-year graduate and post-graduate students of the Madrid Open University, UDIMA. For their final assignment, students are required to carry out a system development project. Projects from two different subjects from different study programs were considered, from the academic year of 2013-14 until 2018-19. In all cases, the projects were supervised by the same tutor.

TABLE 1. List of projects analyzed.

#P	Title	Programme of study	Subject	Year
1	Prototype of a Workflow for Storing Weather Data using Big Data Frameworks	MSA[a]	EMP[c]	13/14
2	User-activity analysis in a Social Telecare Service through Big Data	MSA	EMP	15/16
3	NoSQL Databases: Designing an application prototype based on MongoDB	DCE[b]	EDP[d]	15/16
4	Big Data: Application and Use in Food Systems	MSA	EMP	15/16
5	Semi-structured Data Analysis using Big Data and Map Reduce	MSA	EMP	15/16
6	NoSQL Databases: MongoDB and Apache Cassandra-based prototype development	MSA	EMP	15/16
7	Big Data System for Improved Agricultural Output in Castilla y León	DCE	EDP	18/19
8	Extending a Big Data system to improve agricultural output by predicting the requirement for treated water in countries with scarce water resources.	DCE	EDP	18/19

[a]MSA = University Master's Degree in Software Architecture; [b]DCE = Bachelor's Degree in Computer Engineering; [c]EMP = End-of-Master's Degree Project; [d]EDP = End-of-Degree Project

Table 1 lists the Big Data projects collected for this study [22]–[29], specifying, for each project, its identifier (#P), title, the program of study, the subject that the project is a part of, and the academic year of completion.

In all the cases, the tutor holds a meeting with the student at the beginning of the subject to define the topic (the domain of application) of the Big Data project. In some cases, the topic is directly proposed by the tutor, who also provides the dataset to the student. However, in other cases, the student knows about a dataset of interest and they propose the project topic to the tutor. In any case, the tutor makes a series of recommendations to the students and provides him or her with a summary of the same and some other information (interest, duration, expected results, and so on) that is included in Appendix A, in order to comply with the current University regulations.

Permission was granted by the authors of the projects listed in Table 1 and they were asked to fill in a questionnaire with the information required to complete this research. In all cases, the students willingly and diligently filled in the questionnaires. The questionnaire was validated by a group of experts as we will describe in Section V of this manuscript.

The structure of the questionnaire sent to the students is included in Appendix B of this document. The goal of the data collection and general instructions on how to fill in the requested information was included in the questionnaire. Subsequently, it was divided into a first part with general questions on the characteristics of the projects (type and size of data, system interactivity, project phases, etc.), and the responses to these questions are the basis of Section 3 of this article. The second part of the questionnaire included questions that the students were asked to respond to according to their experience after completing the project and their responses are used to answer the Research Questions of this article. This second part of the questionnaire includes questions about the effort dedicated, the risk and difficulties found, the available resources, recommendations and challenges to face.

After completing the survey, the students submitted their responses, which were organized and then analyzed in order to address all the research questions proposed in this research.

Fig. 1 depicts the main steps of the methodology we followed in this research. The first one consisted on collecting information about past Big Data projects supervised by the authors and selecting them; after that we collected students' permission to work with their project; then we defined and validated the survey; after its validation, the survey was handed to the students who filled it and send it back to the authors of this paper. After that, we organized and analyzed data and, based on that, we finally presented a series of lessons learned from this study.

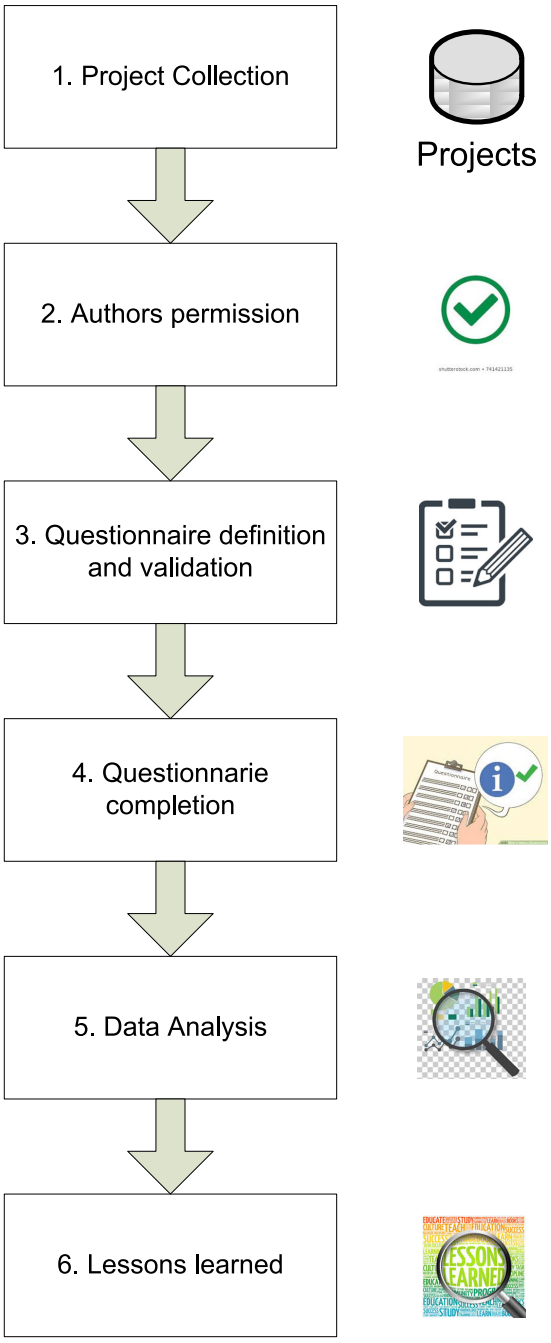


FIGURE 1. Graphical description of the methodology used.

III. DESCRIPTION OF THE PROJECTS

As mentioned in the earlier section, a total of eight projects have been included in this research, from many different fields. Table 2 provides a summary of the projects considered in this research and their keywords, obtained mostly literally from the reports written by the students themselves. This gives the reader an idea of the characteristics of each project.

Apart from their functionality, there are other interesting characteristics of these projects that are worth noting, such as the stages implemented (C = Collecting; S = Storage; P = Preparation; A = Analysis; V = Visualization), the dynamic (or static) nature of the data, their temporal characteristics, size, the interactive nature of the developed system and ad-hoc tools developed as part of the project. All this data is included in Table 3.

As can be seen in Table 3, there is a certain variety in the projects under study, although all of them are focused on using Big Data to solve a problem in a certain business area. Given its prototypical nature (possibly the application that possesses the most characteristics of Big Data analysis), two

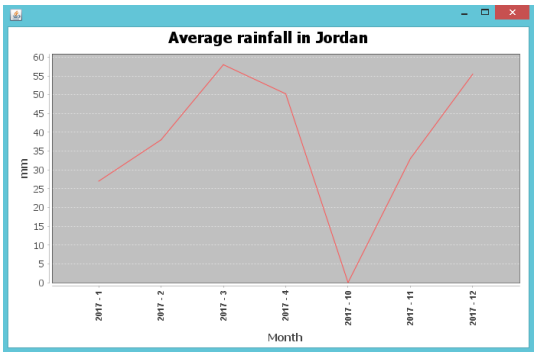
TABLE 2. Projects summary.

#P	Abstract	Keywords
1	Weather systems like the World Meteorological Organization´s Global Information System need to store different kinds of images, data and files. Big Data and its 3V paradigm can provide a suitable solution to solve this problem. This thesis presents some concepts around the Hadoop framework, de facto standard implementation of Big Data, and how to store semi-structured data generated by Automatic Weather Stations using this framework. Finally, a formal method to generate weather reports using Hadoop´s ecosystem frameworks is presented.	Big Data, Hadoop, HDFS, WMO, Global Observing System, IoT
2	Nowadays, the use of Big Data is taken a strength and a very important relevance; in the biggest companies of social sector and service sector are using Big Data technologies that allow to store and treat all the information that they have of users and, in a second way, the incorporate the knowledge of the treatment of this information in the life of the users. In Telecare, with the IP technology in Telecare Unit, the companies will start to use these technologies to store all the information that the unit will send to the control center. With all this information, the companies will be able to discover patterns of user's behavior, detect some illnesses like, for example, Alzheimer; but the most important action that the companies will be able to have more information related to the situation of all devices and sensors installed in user's home when the emergency alarm is raised. A work of data mining in telecare calls using Big Data will be presented.	Big Data, Hadoop, Map Reduce, Mahout, Data mining, Telecare Users, Calls, Resources optimization
3	This project aims to give an overview of databases, focusing in nosql systems, especially in the open source mongodb tool. In the last years the field of databases has evolved to a nosql trend which is providing new capabilities to the traditional models. Besides, it has Been adapted to current society changes. I decided to use mongodb to perform this model because currently it is one of the most used open source systems in the business environment. The application to be developed will be a nosql database using mongodb. The case study will be a medical centre in which patients will be represented as documents in order to ease the search process carried out by either doctors and nurses and administrative personnel. When using this management system the user will be able to customize Every document for Each patient's specific features. Moreover, the user will be able to search amongst the information introduces in a quick and efficient manner.	Not available
4	The present work studies the role of nutrition in the world through the analysis of large volumes of data related to food products consumption in different parts of the world, with the aim of finding out the eating habits globally and thus provide objective, reliable and easy to understand information, that help countries and its governments to create programmed and policies that allow them to address the nutritional problems and improve health conditions. The result of this work shows a set of nutrition indicators obtained by analyzing over eighty thousand food labels from food products from an open database, using tools such as *Big Data, Open Data y Data Mining.* This publication will serve as an introduction to nutrition problems and intends to make a contribution on the topic, raising awareness of how diet has a direct influence in the quality of life of human beings. This work will be particularly useful for organisms and healthcare professionals, as well as the general public.	Nutrition, Healthy Food, Big Data, Open Data, Collaborative Data, Diet, Nutritional Indicators
5	The purpose of this work is to apply big data analytic techniques to study an internet web forum data set. The subject of the forum in this case is technology and programming languages, but the implementation of this work is applicable to any forum regardless the subject as long as it has a specific data structure. Basically we will focus on Hadoop framework that we will be using as tool that provides us with techniques and infrastructure. Regarding the techniques we will mostly use map reduce, and regarding the infrastructure Hadoop provides us with a cluster implementation that allows us to use big data techniques and parallel data processing. The idea is to build a system which is able to analyze the forum's data using the map reduce technique, and in order to do so different consultations will be written to answer these kind of concerns about the user's data. Basically this tool will answer those concrete questions like how many users know a specific technology or which technologies are the most used ones in a specific time frame.	Forum Technology, Knowledge discovery, Big data,Map reduce
6	This document aims to give an overview of the NoSQL databases, explaining the origin and reasons of its use, mentioning some use cases in the market, and detailing their characteristics. The NoSQL aggregate data model enables to fragment and distribute the data across multiple nodes (servers) using a transactional control of basic availability, self-state, and eventual consistency (BASE) that works hand to hand with CAP theorem: consistency, availability, partition tolerance. The NoSQL databases design allows to have a boost in performance versus relational databases, minimizing bottlenecks, maximizing the use of RAM and hard drives, and finally, using consistent data hashing. Among the different types of NoSQL storage, we have: column-oriented, document, key-value, graphs, or multistorage. MongoDB is a document type database and Cassandra, column-oriented. Both are mature enough and are widely used. The main objective to resolve in present work is to develop a prototype to demonstrate the use and benefits of NoSQL databases, MongoDB and Apache Cassandra.	DDBB, NoSQL, MongoDB, Apache Cassandra, Prototype, Server. API, C++
7	At present, huge number data are generated within any area of knowledge and to be able to take advantage of them new tools have been developed. The ultimate objective of Big Data is to make decisions that improve the results of the organization. The objective of this dissertation is the development and implementation of a big data system that improves agricultural performance in Castilla y León. This system allows to load the data, process them, visualize them and obtain models capable of predicting the amount of precipitations that will fall next year. This knowledge will allows taking decisions about the type of crop that will be more profitable for the farmer.	Big data, ,Crops, Rainfall, Data processing, Data mining, Time series
8	This work is the continuation of the Big data system that improves agricultural performance in Castilla y León. The project aims to transform the application that was made with a particular purpose into a more universal solution that would cover all the regions of the world as well as different meteorological variables. Furthermore, new data mining algorithms are developed, and their performance is contrasted in different scenarios. Although the system can be useful in any geographical area, the tests focus on the weather stations of the Near East countries, especially Jordan. This region is characterized by water shortage and a system for rainfall and temperature forecast would be especially interesting in this environment. The possibility to predict weather condition for the next year would allow the public administration to estimate the needs of alternative water resources. Additionally, an important part of this work is dedicated to the problem of data mining performance with a high level of missing values. Alternative algorithms, apart from time series, were developed to improve the accuracy of predictions. This project represents a great advance in regard to the original system.	Big Data, Data Mining, NoSQL Meteorology, Near East

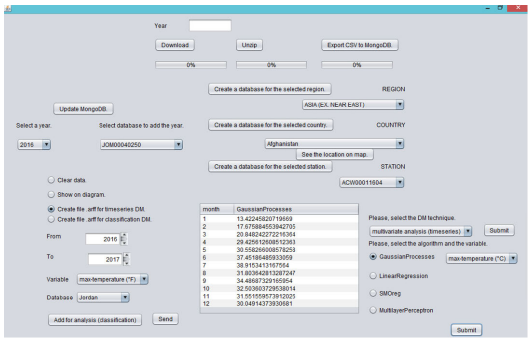
TABLE 3. Projects main information.

#P	Stage C	S	P	A	V	Dynamic data (not static)	Time series	Data size	Interactive System	Ad-hoc tools
1		√	√				√	8 MB[1]	√	√
2	√		√	√	√	√		30 GB		√
3	√	√	√	√	√	√		100 MB	√	
4		√	√	√	√			20 GB	√	√
5	√		√	√			√	15 GB	√	√
6	√	√			√	√		100 MB		
7	√	√	√	√	√	√	√	100 GB	√	√
8		√	√	√	√	√	√	20 GB	√	√

[1]In some cases, the data size is unusual in Big Data as they are small prototypes.



a) Past rainfall visualization



b) Temperature forecasting

FIGURE 2. Screenshots of some of the big data systems.

screenshots of some of the projects are included in Fig. 2 as an example of the applicability of this type of projects, displaying some of the results obtained (visualization of historical rainfall data in Fig. 2.a and predicting next values in a meteorological time series of max temperature in Fig. 2.b).

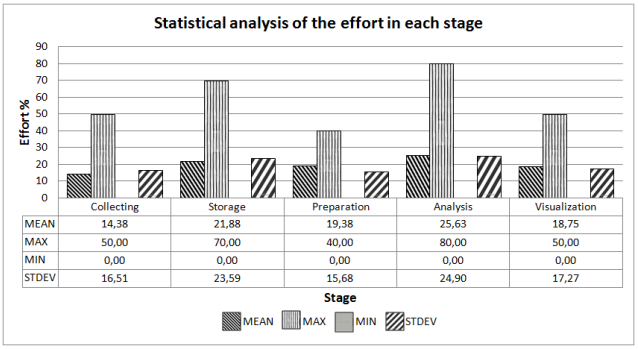
IV. RESULTS AND DISCUSSION

After collecting the data given by the authors of these projects through the aforementioned questionnaire, the results obtained were analyzed in order to respond to each of the seven research questions posed in this work.

A. RQ1 - WHAT PERCENTAGE OF EFFORT IN A BIG DATA PROJECT IS REQUIRED FOR EACH STAGE?

The goal of the first question is to learn about the stages of a Big Data project that require the most work. In spite of the slight limitation that some projects do not include certain stages, a statistical analysis was made in order to determine the average values and other statistical values in each stage. We should clarify that in this RQ1 we did not analyze the aspects that make it necessary to spend more time to complete some stages (those aspects, such as low data quality, will be discussed later).

The results obtained are shown in Fig. 3.



	Collecting	Storage	Preparation	Analysis	Visualization
MEAN	14,38	21,88	19,38	25,63	18,75
MAX	50,00	70,00	40,00	80,00	50,00
MIN	0,00	0,00	0,00	0,00	0,00
STDEV	16,51	23,59	15,68	24,90	17,27

FIGURE 3. Statistical analysis of the effort for each stage.

As may be seen, on average, the stage that requires the most effort is Analysis, followed by Storage. In third place we have Preparation, closely followed by Visualization. Finally, the stage that involves the least effort is Collection.

If we analyze the maximum values for each stage, Analysis and Storage still occupy the highest positions. These data can serve as a reference when planning Big Data projects, keeping in mind that, based on the experience described in this research, the two aforementioned stages appear to require the greatest effort.

Nevertheless, it must also be remembered that there were projects that did not involve significant effort in said stages owing to their particular characteristics. This, linked to the wide variability in efforts in all stages but mainly in Analysis and Storage, urges caution when planning this type of projects by making a prior analysis of the stages that they will consist of.

B. RQ2 - WHAT ARE THE MAIN RISKS/DIFFICULTIES THAT MAY THREATEN A PROJECT?

In contrast to the previous question, this is of a more qualitative nature and therefore, requires a detailed analysis of

the responses given by each participant. The goal is to find difficulties that are common to the largest number of projects possible, so they may serve as a precautionary measure for future projects.

In this case, the results of the analysis are displayed in Table 4, which lists the risk detected, the number of projects that have this risk (column arranged by number of projects), and the average impact (values between 1 and 5, where 5 stands for the greatest impact).

TABLE 4. Most-used technologies.

Risk	Number of Projects	Average impact
Low data quality	5	4
High initial learning curve	4	4
Need to learn about the domain of application	3	4.7
Difficulty in finding the required infrastructure for Big Data	2	4

In the table above, only risks found in two or more projects have been included, as being more frequent and therefore more representative. Of the results obtained, low data quality appears to be the most recurring problem (in 5 out of 8 projects), and it must be considered as a risk factor in any big data project. A prospective analysis in search of data of sufficient quality is an important recommendation to take into account before fully commencing a Big Data project.

The second most frequent inconvenience (4 out of 8 projects) is the high initial learning curve, which may significantly slow down project development. The third most frequent aspect (3 out of 8 projects) is the importance of learning about the domain of application. It also has the greatest impact value (4.7). It is therefore essential to have experts in the domain of application in order to successfully execute projects of this type. Finally, the difficulty of finding a big data infrastructure at the academic level was also pointed out in 2 of the 8 projects.

C. RQ3 - WHAT IS THE DEGREE OF DIFFICULTY OF EACH STAGE IN RELATION TO THE OTHER STAGES?

Although the earlier question already deals with some difficulties, many of these do not belong to any concrete stage but are general problems. Therefore, it is difficult to arrange the difficulty of each stage.

In order to answer this question, RQ3 calculates an average of the difficulties allotted to each stage by the interviewees (values between 1 and 5, where 5 stands for the highest difficulty). The results are displayed in Table 5 according to the total value of each stage.

According to the results obtained, the most difficult stage is the Data Analysis stage, followed by Data Preparation. These results are consistent with the findings of other researchers in the field of data mining, a pattern that also appears in this type of Big Data projects.

TABLE 5. Relative difficulty of each stage.

Order of difficulty	Stage	Average difficulty
1	Data Analysis	1.83
2	Data Preparation	2.43
3	NoSQL storage	2.71
-	Data Visualisation	2.71
5	Data Collection	3.33

The third-most difficult stage is jointly occupied by Storage and Visualization, which are more characteristic of this type of projects. Finally, Data Collection appears to be the least difficult stage.

D. RQ4 - WHAT ARE THE MOST FREQUENTLY USED TECHNOLOGIES IN EACH STAGE?

The goal of this research question is to learn about the technologies that are most used in each stage of a Big Data project. For this, each interviewee was asked to note down the tools/technologies (not directly implemented by them) used in each stage. The results are displayed in Table 6 where the most-used technologies are grouped by stage and, in every case, the project percentage (how many of the total 8?) that utilized them.

TABLE 6. Main risks encountered.

Stage	Technology	%Use
Data Collection	MS SQL Server	12.5% (1/8)
	Amazon Web Services	12.5% (1/8)
NoSQL Storage	MongoDB	50% (4/8)
	HDFS	25% (2/8)
	Apache Cassandra	12.5% (1/8)
	Robo3T	12.5% (1/8)
	Studio3T	12.5% (1/8)
Data Preparation	MS SQL Server	12.5% (1/8)
	MongoDB	12.5% (1/8)
	Apache Spark	12.5% (1/8)
	Robo3T	12.5% (1/8)
	Studio3T	12.5% (1/8)
	Pig	12.5% (1/8)
Data Analysis	Hadoop	25% (2/8)
	Mahout	12.5% (1/8)
	Apache Spark	12.5% (1/8)
	Weka	12.5% (1/8)
	Time series forecasting libraries	12.5% (1/8)
Data Visualization	JFreeChart	25% (2/8)
	Apache Zeppelin	12.5% (1/8)
	Hue	12.5% (1/8)

The above table shows that the most-used technology in the Storage stage is No SQL and Hadoop in the Data Analysis stage. At the same time, the JFreeChart library is the most-used tool in the Visualization stage. In the Data Preparation stage, despite using multiple tools, no tool is repeated in more than one project. The same occurs in the Data Collection stage.

E. RQ5 - WHAT IS THE STAGE THAT HAS THE LARGEST AMOUNT OF AVAILABLE RESOURCES, TOOLS, LITERATURE, ETC. AND WHAT IS THE STAGE THAT HAS THE LEAST?

This question again seeks to arrange the stages, here, based on the number of existing resources available to the teams that carry out Big Data projects. For this, the mean of the values awarded by the interviewees has been calculated, in order to determine which stages have greater resources and which stages have fewer.

The results in Table 7 are arranged from more to less available resources, according to the average value awarded by the interviewees (values from 1 to 5, where 5 stands for the least number of resources).

TABLE 7. Availability of resources in each stage.

Order amount of available resources	Stage	Average value available resources
1	NoSQL storage	1.86
2	Data Collection	2.33
3	Data Preparation	2.57
4	Data Analysis	3
5	Data Visualization	3.29

The results of the previous table indicate that the stages of Storage and Data Collection, which are very characteristic of Big Data projects, are those that have the most resources. Nevertheless, stages that deal with traditional data analysis such as Preparation and Data Analysis itself do not have as many resources. This points to the need for greater research by the scientific community on developing and/or adapting existing pre-processing and data analysis techniques to large quantities of data.

The Visualization stage is also inherent to Big Data, and has the least number of available resources, thus making it another important research avenue regarding the development of techniques/tools for visualizing large quantities of data.

F. RQ6 - WHAT ASPECTS (RECOMMENDATIONS) ARE KEY TO ACHIEVING A SUCCESSFUL BIG DATA PROJECT?

Again, this question is of a more qualitative nature and therefore, requires a detailed analysis of the responses given by each participant. The goal is to find common recommendations in the largest number of projects possible, so they may serve as aspects to be considered in future projects.

In this case, after analysis, the results are displayed in Table 8, which lists the recommendation detected, the number of projects where it appears (column arranged by number of projects), and its average relevance (values greater than or equal to 1 where 1 stands for significant relevance).

Only recommendations that appear in two or more projects have been included in the previous table, for purposes of representation, although there were many others that have not been included for being too specific.

TABLE 8. Key recommendations for success.

Recommendations	Number of Projects	Average relevance
To know and select the correct technologies and tools	6	2
Access complete and high quality datasets	3	1.3
Have big data infrastructure	3	4.3
Comprehend the Big Data ecosystem	2	2.5
Knowledge of Big Data languages	2	2.5
Offer a good data visualization	2	3.5

TABLE 9. Most-used technologies.

Challenges	Number of Projects	Average relevance
Having a greater number of open and high-quality datasets	3	2
Standardizing the term Big Data	2	1
To have mechanisms to guarantee data security	2	1.5
New techniques for the analysis of non-conventional data	2	2.5
New NoSQL tools	1	1
To transfer it to new disciplines	1	1
Define data quality standards	1	2
To have techniques to measure information quality	1	3
To have training plans in this area	1	3
To have an ethical framework for data processing	1	5

The conclusions that may be drawn from the table is that it is essential (in 6 out of 8 projects) to have knowledge of the necessary technologies and tools to undertake big data projects, as this recommendation has a relatively high relevance. Therefore, it is essential to have a training plan for this type of project.

The second-most frequent recommendation (3 out of 8) is to ensure the availability of a complete and high quality dataset, with an impact value close to the maximum value of 1. It was also recommended in 3 projects that a good Big Data infrastructure be available from the beginning, although it had a significantly lower impact (4.3).

Other recommendations appear somewhat less frequently (2 out of 8 projects) although they have a certain relevance (between 2.5 and 3.5), of which a correct understanding of the Big Data ecosystem and knowledge of Big Data languages are worth mentioning. Again, training or prior experience are shown to be essential requirements for successfully executing this type of project,

Finally, providing good data visualization is a success factor in 2 out of 8 projects, therefore, boosting this area would seem to be good practice.

G. RQ7 - WHAT EXISTING PROBLEMS/CHALLENGES MUST BE FACED BY THE BIG DATA COMMUNITY?

Finally, the survey participants responded to a question on what, based on their experience and judgment, are the greatest challenges faced by the scientific community in the field of

TABLE 10. Expert responses.

	Element to assess	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6	Expert 7	Expert 8	Expert 9	Expert 10	Average	V Aiken
Method	**M1.** How far do you consider that the problem addressed is worth researching?	4	3	3	4	3	4	3	3	4	4	3.5	.875
	M2. How far do you consider that the RQs are well defined?	3	4	3	4	3	4	2	3	3	4	3.3	.825
	M3. How far do you consider that the methodology used is appropriate to answer the RQs?	4	4	4	3	4	3	3	4	3	3	3.5	.875
	M4. How far do you think that the number and characteristics of the analyzed projects is enough to obtain interesting conclusions?	3	3	4	2	3	3	3	2	3	3	2.9	.725
	M5. How far do you think that the results obtained are generalizable to Big Data projects overall (not in the academia)?	2	2	3	1	3	1	2	1	1	2	1.8	.45
	M6. How far do you think that the results obtained are directly applicable?	3	2	3	4	3	3	2	2	1	3	2.6	.65
Questionnaire	**Q1.** How far do you think that Item 8 in our questionnaire is convenient?	4	3	3	3	4	4	3	3	4	3	3.4	.85
	Q2. How far do you think that Item 9 in our questionnaire is convenient?	2	4	4	4	3	3	4	3	4	2	3.3	.825
	Q3. How far do you think that Item 10 in our questionnaire is convenient?	4	4	4	3	4	3	4	4	4	4	3.8	.95
	Q4. How far do you think that Item 11 in our questionnaire is convenient?	4	4	4	4	4	3	2	5	4	3	3.7	.925
	Q5. How far do you think that Item 12 in our questionnaire is convenient?	2	4	3	4	3	4	4	4	4	4	3.6	.9
	Q6. How far do you think that Item 13 in our questionnaire is convenient?	4	3	4	4	4	4	3	4	2	4	3.6	.9
	Q7. How far do you think that Item 14 in our questionnaire is convenient?	4	4	1	4	4	3	4	4	3	4	3.5	.875

Big Data. The goal, once again, is to find challenges that are common to multiple participants.

The results are displayed in Table 9, which lists the challenge detected, the number of projects where it appears (column arranged by number of projects), and its average relevance (values greater than or equal to 1 where 1 stands for significant relevance). In this case, as there was not an excessive number of challenges and given the interest regarding the ideas provided as possible lines of future research, all responses made by the interviewees have been presented.

Of the results obtained, the major challenge faced by the community appears to be the difficulty in finding open and high quality datasets to work with. In this regard, an effort must be made to provide the community with the datasets to be worked with in each research project.

Another challenge that was surprising but merits attention is: standardization of the term Big Data, frequently used on a global level from different perspectives. To have mechanisms that guarantee the security of the data and to have techniques to analyze large quantities of unconventional data was also a recurring challenge.

Finally, there are also other challenges such as NoSQL tools that do not possess the limitations of the current ones, taking big data to new unexplored disciplines, defining data quality standards and techniques to measure said quality, as well as having access to training plans in an area and to have an ethical global framework for processing big data.

V. EXPERT VALIDATION

We have implemented a procedure in order to assess the validity of our method overall and the questionnaire used (Appendix A) in particular. To do so, we count on a group of 10 independent experts in the Big Data area. All those experts have experience in teaching Big Data courses and supervising the development of academic Big Data projects.

First, we provided the experts with information about the method used in our research and the questionnaire itself. After that, we asked them to answer a series of questions (denoted M1-M6 in Table 10) about the validity and applicability of our method and the results obtained, and we also asked them to rate the convenience of the items used in our questionnaire denoted Q1-Q7 in Table 10 (only those related to the Research Questions). In both cases, experts were required to assess each element with a value in the set {0,1,2,3,4} where 4 represents the most positive possible feedback.

In order to summarize the global value given by the expert for each assessed element, we decide to use the *V Aiken* statistic, a commonly used approach to summarize research relevance ratings obtained from experts. We decided to use this approach since it has demonstrated to be useful in the educational area for assessing research similar to the one presented in this paper [30], [31]. The formula used to calculate this statistic is defined in (1).

$$V = \frac{\sum S_i}{n(c-1)} \tag{1}$$

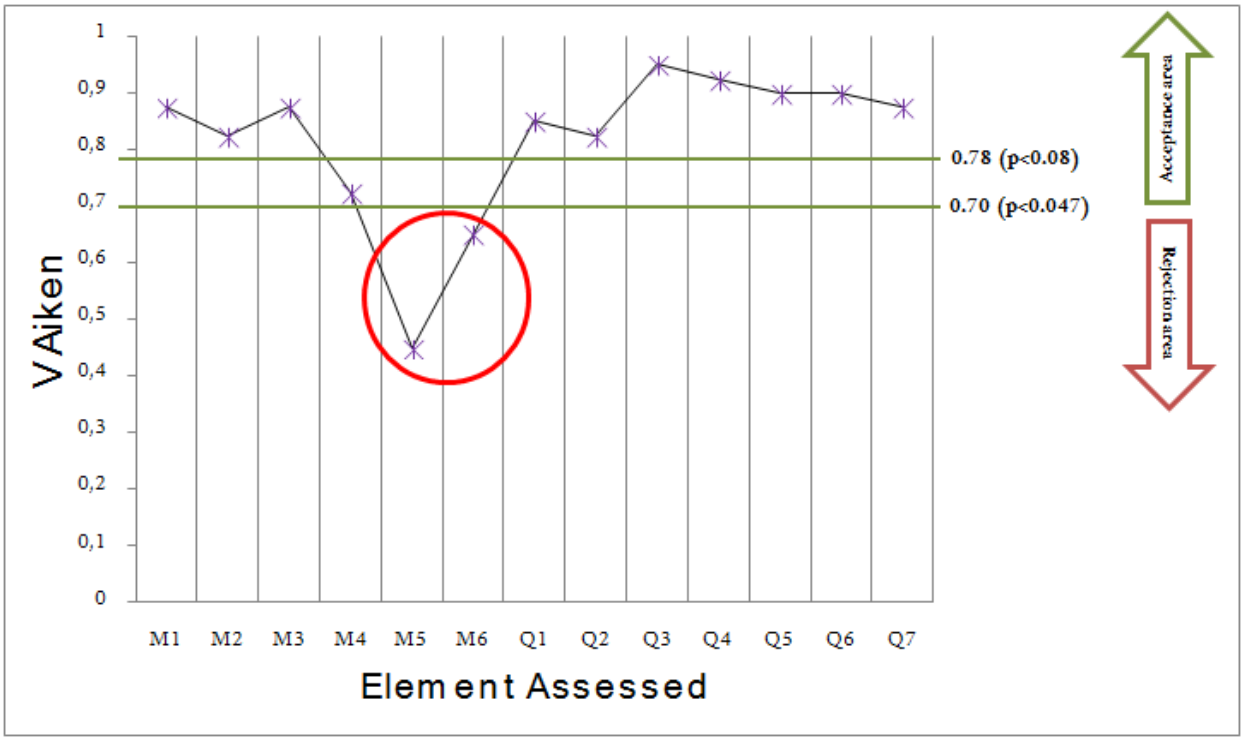


FIGURE 4. Graphical summary of V Aiken expert statistical analysis.

where S_i represents the sum of the values provided by expert to each assessed element, n is the number of experts (10 in this case), and c is the number or possible categories to rate (5 in this case).

Considering the results obtained (Table 10), we can conclude that, according to experts, the problem addressed in this paper is worth researching, the method is appropriate, and the research questions are well defined with a statistical significance level of $p<0.008$. In addition we can also conclude that the projects considered are enough for this research at a significance level of $p<0.047$. With respect to the direct applicability of the results, the V Aiken indicator stays in 0.65, which means that the assessment of this item is really close (but below) to the statistical significance limit of 0.7 (obtained in V Aiken right-tail probability table). Finally, we cannot conclude that the results obtained can be transferred out of academic field. Regarding the questionnaire items, we can conclude that all of them are convenient at a significance level of $p<0.008$. Fig. 4 graphically summarize all these statistical findings.

VI. CONCLUSION AND FUTURE LINES OF RESEARCH

The purpose of this research is to derive the lessons learned from the development of Big Data projects in the academic field. For this, a survey was carried out among different students who have carried out this type of project in end-degree subjects related to Computer Engineering. As such, it is a low-level study whose preliminary results are meant only as an initial source of knowledge on the development of these projects in academic environments. From the expert assessment, we understand that it would be necessary to consider and analyze a larger number of similar projects in order to confirm the preliminary findings of this research at a higher statistical confidence level.

The findings obtained in this research can be useful to stakeholders (students, professionals, lecturers and even institutions) who encounter similar projects. This is the primary contribution of this study and its most important findings are summarized below:

1) Regarding the planning and procurement of resources for Big Data projects, it must be noted that the stages of Data Analysis and Storage require the most effort, whereas Visualization and Collection require the least. Of course, it is essential to examine the type of project to be undertaken and the above-mentioned conclusions may be altered due to the specific nature of the project.
2) The main risks of these projects are low data quality and high initial learning curve, apart from the need to learn about the domain of application.
3) The most difficult stage is usually the Data Analysis stage, whereas the Collection stage is usually the least difficult.
4) The data storage technology most widely used in these projects is MongoDB, followed by HDFS. Hadoop and Apache products are also extensively used.
5) There are different resources available to the community for Big Data projects but based on the stage under

consideration, they may vary significantly. Data Storage has the largest amount of resources and Visualization the least.

6) Before undertaking a project of this type, it is essential to have knowledge of the technologies and tools, to have the proper technical infrastructure and to have complete and high-quality datasets.
7) Despite the significant progress made by the scientific community in recent years, there are still unsolved problems in the area of Big Data (not necessarily academic), which provide us with important research opportunities for the future.

The above-mentioned findings are supported by the positive results obtained during the expert validation process conducted as a part of our research, which can be considered as one of the strongest point of our study. Regarding the applicability of the obtained results, we would like to highlight that the lessons learned from this research can be useful mainly in similar academic scenarios. However, some of the findings obtained might be also useful for the Big Data community, not only limited to the academia. But to confirm this point it would be necessary to conduct a wider research out of academia as suggested by experts in their assessment.

Finally, the research questions analysis and the conducted research in general have helped us to establish a series of lines of research in the area of academic Big Data project development as well as in the area of Big Data itself. As suggested by some of the experts involved in this research, the direct applicability of the results obtained is not an easy task. Therefore, an initial future line of research would be the definition of a formal guide or process model, similar to CRISP-DM in Data Mining area [32], which can incorporate the lessons learned from this research. It would also be interesting to apply that model in big data courses and check its impact in students' performance and dropout rate. A new exhaustive research must be conducted with experiments to confirm the good results obtained in this paper. Another line of research could be the proposal of methodologies, or the modification of existing methodologies for Big Data projects that would consider the average percentage of effort required for each stage. The knowledge of these values could aid in better time and resource management.

Some other future lines of works can be found below:

- The community must strive to provide other researchers with datasets, provided they are of sufficient size and quality.
- The stages of Data Analysis and Preparation continue to present the most difficulties for people who take up this type of project. More research into these stages would provide new and improved resources to deal with them.
- Researchers must work in order to provide more resources to the community (documentation, tools, techniques, etc.) primarily in the stage of Data Visualization as well as the Data Analysis stage which, although is of long-standing in the field of data mining, requires important adaptations for Big Data problems.
- The community must re-evaluate the current meaning of the term Big Data and establish a standard definition and procedures that help to maintain its essence and distinguish it from other current concepts and proposals.
- Mechanisms to guarantee data security are required, as well as to define their quality and to ensure ethical behavior and use.
- It is also necessary to define techniques and tools for large quantities of unconventional data, in addition to new tools for efficient storage of said data.
- Today there are clear fields of application of Big Data, such as education or health. Other less-explored domains such as the environment, for example, could be suitable for future research.

APPENDIXES

APPENDIX A

Tutor guidelines for project execution

Bachelor's Degree in Computer Engineering

End-of-Degree Project Proposal
Academic Year-Semester**2018-2019 - 2nd Semester**
Project proposed by

Lecturer

Name of the lecturer: <lecturer's name>

Proposal

Project Title: ***Developing a knowledge discovery system under Big Data approach***

Number of proposal: 1

Number of students: 1+

1- Summary: (include a maximum of 200 words)

From its very onset, managing large quantities of data has been one of the primary challenges in the area of data mining. Possessing a large quantity of records and an elevated number of characteristics or attributes for each record has been an important challenge in this discipline.

Although this is a pre-existing issue and there is specific terminology to refer to it, in recent years, a term that encompasses the most relevant ideas on processing large quantities of data has emerged. This term, Big Data, refers to obtaining, storing, processing, viewing and analyzing large quantities of data.

The primary goal is to design, implement and evaluate a computer system that resolves any issue related to Big Data. In an ideal scenario, the system could be responsible

for obtaining, storing, processing, visualizing and analyzing large quantities of data, however it will also admit projects that only focus on some of these tasks if the difficulty level of the problem solving requires it. It will not admit solutions that require the mere application of already existing tools, save when the necessary integration requires a strong IT development.

2. Interest in the project and potential beneficiaries: (include a maximum of 100 words)

The project must solve, using Big Data techniques, a business-related problem within the scope of the analyzed data, or a more fundamental problem within the discipline of Big Data, making clear the benefits to potential clients/users of the domain of the analyzed data. The student may take data from an area close to them (business, for example), or seek an open data source, provided they conform to the characteristics of Big Data (volume, speed of updates, variety).

3-Duration and depth of work: (include a maximum of 100 words)

In order to accomplish their end-degree project, the student must work independently, thoughtfully, and with enough technical depth, in order to analyze existing needs within the scope of study and propose an appropriate solution. The task should take approximately 300 hours (including the drafting of the final report and the presentation of the end-degree project).

The end-degree project report must include information on the expenses and duration (planning) of the project, both in the execution of the end-degree project and in its possible extrapolation to a wider and more realistic environment.

4-Profile of the lecturer supervising this project: (include a maximum of 50 words)

<lecturer's name> is Lecturer of Knowledge Extraction at this University. He holds a doctorate in Computer Engineering and his thesis dealt with the analysis of structurally complex data, including time series. He is the author of more than fifteen articles of impact.

APPENDIX B

QUESTIONNAIRE – BIG DATA PROJECTS

The goal of this questionnaire is to compile information on your experience in developing a Big Data project. With this information, we hope to carry out research that will help us to draw conclusions on the time required to execute a Big Data project, the difficulties encountered, the lessons learnt, etc. Your answers must be based on your experience of the project.

Initially, it is assumed that all Big Data projects include the stages of: **Data Collection, NoSQL Storage, Data Preparation, Data Analysis, and Data Visualization.** Nevertheless, it is possible that your project did not include some of these stages. This is not an impediment to filling this questionnaire.

By filling in this questionnaire you consent to the sharing of your replies, the conclusions drawn from them, as well as a general description of your project, with the scientific community. Under no circumstances, will your personal data or data regarding your academic performance ever be disclosed.

GENERAL PROJECT DATA

Project Title or Topic:

1. Indicate the stages included in your project

Stage	Yes	No
Data Collection		
NoSQL Storage		
Data Preparation		
Data Analysis		
Data Visualization		

2. Was the data that you worked with static (fixed, without updates over time) or dynamic (permits the addition of new data to existing data)?

3. Did you work with time series data (indicate yes or no)?

4. Indicate the approximate size of the data that you worked with (e.g., 20 Gigabytes).

5. Indicate whether your system permitted end-user interaction or if it was a non-interactive system.

6. Did you develop any custom tool within the system? (answer yes or no)

7. If your answer to Question 6 was yes, indicate, for each developed tool, the stage(s) where the tool was used, and the language used to implement it.

Tool	Stage(s)	Language
...		

LESSONS LEARNED

8. Indicate the percentage of effort dedicated to each stage of the project (the total must be 100, if you did not implement a certain stage, mark it with 0)

Stage	%Effort
Data Collection	
NoSQL Storage	
Data Preparation	
Data Analysis	
Data Visualization	

9. List the principal risks/difficulties that, in your experience, may endanger the success of a Big Data project. For each risk, indicate its degree of impact on the project from 1 to 5 (1 = low risk; 5 = highest possible risk), and the stages where this risk was encountered (you may indicate all risks that you consider to be important, even if you did not encounter them in your project)

Detected Risk	Impact	Stage(s)
...		

10. Mark with consecutive numbers (1, 2, ...) each stage of your project according to the relative degree of difficulty with regard to the other stages (1 indicates the stage with the highest difficulty; if you did not execute a certain stage, leave it blank and move to the next stage)

Stage	Difficulty
Data Collection	
NoSQL Storage	
Data Preparation	
Data Analysis	
Data Visualization	

11. List the (existing, not created by you) Big Data technologies used in each stage of your project (if you did not execute a certain stage, leave it blank; if you did not use any existing tool in a certain stage, indicate with a hyphen)

Stage	Tools
Data Collection	
NoSQL Storage	
Data Preparation	
Data Analysis	
Data Visualization	

12. From your experience, **rank** with consecutive numbers (1, 2, ...) each stage of your project, based on the relative quantity of publicly available resources (tools, books, articles, forums, manuals ...) in each stage, with regard to the other stages (1 stands for the stage with the most available resources; if you did not execute a certain stage, leave it blank and move to the next stage)

Stage	Resources
Data Collection	
NoSQL Storage	
Data Preparation	
Data Analysis	
Data Visualization	

13. In your experience, what aspects are essential (by way of recommendations) to successfully execute a Big Data project? (Note them in order of importance, where 1 = Most important)

Relevance	Recommendation
1	
2	
3	
4	
5	
...	

14. In your experience, what are the main challenges (existing problems that are worth exploring) that must be faced by the scientific community in the field of Big Data? (Note them in order of importance, where 1 = Most important)

Relevance	Challenges
1	
2	
3	
4	
5	
...	

ACKNOWLEDGMENT

This paper was drafted as part of Juan A. Lara's research stay during 2019-2020 at Jordan University of Science and Technology, JUST (Jordan), which partially sponsored this research. The authors would like to thank JUST's and MEU's experts on Big Data for their valuable advice on how to design the data collection stage and their support in validating the methodology and survey used. The authors gratefully acknowledge the support of the students who executed the projects mentioned in this paper and their willingness to complete the survey.

REFERENCES

[1] C. Lynch, "How do your data grow?" *Nature*, vol. 455, no. 7209, pp. 28–29, Sep. 2008.

[2] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery: An overview," in *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. Menlo Park, CA, USA: MIT Press, 1996, pp. 1–34.

[3] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, Apr. 2015, doi: 10.1016/j.ijinfomgt.2014.10.007.

[4] A. De Mauro, M. Greco, and M. Grimaldi, "A formal definition of Big Data based on its essential features," *Library Rev.*, vol. 65, no. 3, pp. 122–135, Apr. 2016, doi: 10.1108/lr-06-2015-0061.

[5] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Inf. Sci.*, vol. 275, pp. 314–347, Aug. 2014, doi: 10.1016/j.ins.2014.01.015.
[6] D. Laney, "3D data management: Controlling data volume, velocity, and variety," *META Group Res. Note*, vol. 6, no. 70, pp. 1–4, 2001.
[7] P. B. Goes, "Design science research in top information systems journals," *MIS Quart., Manage. Inf. Syst.*, vol. 38, no. 1, pp. iii–viii, 2014.
[8] D. Gu, J. Li, X. Li, and C. Liang, "Visualizing the knowledge structure and evolution of big data research in healthcare informatics," *Int. J. Med. Inform.*, vol. 98, pp. 22–32, Feb. 2017, doi: 10.1016/j.ijmedinf.2016.11.006.
[9] H. Liao, M. Tang, L. Luo, C. Li, F. Chiclana, and X.-J. Zeng, "A bibliometric analysis and visualization of medical big data research," *Sustainability*, vol. 10, no. 2, p. 166, Jan. 2018.
[10] S. Bakken and T. A. Koleck, "Big data challenges from a nursing perspective," in *Big Data, Big Challenges: A Healthcare Perspective* (Lecture Notes in Bioengineering). M. Househ, A. Kushniruk, and E. Borycki, Eds. Cham, Switzerland: Springer, 2019.
[11] G. C. Nobre and E. Tavares, "Scientific literature analysis on big data and Internet of Things applications on circular economy: A bibliometric study," *Scientometrics*, vol. 111, no. 1, pp. 463–492, Apr. 2017.
[12] L. Tu, S. Liu, Y. Wang, C. Zhang, and P. Li, "An optimized cluster storage method for real-time big data in Internet of Things," *J. Supercomput.*, to be published, doi: 10.1007/s11227-019-02773-1.
[13] D. Gil, M. Johnsson, H. Mora, and J. Szymanski, "Review of the complexity of managing big data of the Internet of Things," *Complexity*, vol. 2019, pp. 1–12, Feb. 2019, doi: 10.1155/2019/4592902.
[14] A. Mitra and K. Munir, "Big data application in manufacturing industry," in *Encyclopedia Big Data Technologies*, S. Sakr and A. Y. Zomaya, Eds. London, U.K.: Springer, 2019, pp. 1–7.
[15] Y. Lu and X. Xu, "Cloud-based manufacturing equipment and big data analytics to enable on-demand manufacturing services," *Robot. Comput.-Integr. Manuf.*, vol. 57, pp. 92–102, Jun. 2019.
[16] S. Ren, Y. Zhang, Y. Liu, T. Sakao, D. Huisingh, and C. M. Almeida, "A comprehensive review of big data analytics throughout product lifecycle to support sustainable smart manufacturing: A framework, challenges and future research directions," *J. Cleaner Prod.*, vol. 210, pp. 1343–1365, Feb. 2019.
[17] M. Huda, U. Ulfatmi, M. J. Luthfi, K. A. Jasmi, B. Basiron, M. I. Mustari, A. Safar, W. H. Embong, A. M. Mohamad, and A. K. Mohamed, "Adaptive online learning technology: Trends in big data era," in *Diverse Learn Opportunities Through Technology-Based Curriculum Design*, D. Williams and N. Harkness, Eds. 2019, pp. 163–195.
[18] B. K. Daniel, "Big Data and data science: A critical review of issues for educational research," *Brit. J. Educ. Technol.*, vol. 50, no. 1, pp. 101–113, Jan. 2019, doi: 10.1111/bjet.12595.
[19] J. Shamsi, A. Burney, B. H. Butt, and F. Khan, "Teaching big data curriculum: Methods, practices, and lessons," in *Proc. Conf.* At Las Vegas, NV, USA: BDA EdCon, 2014, pp. 1–4.
[20] J. Eckroth, "Teaching future big data analysts: Curriculum and experience report," in *Proc. IEEE Int. Parallel Distrib. Process. Symp. Workshops (IPDPSW)*, Lake Buena Vista, FL, USA, May/Jun. 2017, pp. 346–351.
[21] A. Nagar, "Developing big data curriculum with open source infrastructure," in *Proc. ACM SIGCSE Tech. Symp. Comput. Sci. Educ.*, 2017, pp. 700–701.
[22] M. Almeida, "Prototipo de un flujo de trabajo para el almacenamiento de datos meteorológicos utilizando frameworks para big data," Ph.D. dissertation, Dept. Comput. Sci., Udima, Madrid, Spain, 2014.
[23] A. Moreno, "Análisis de actividad de un servicio de teleasistencia social mediante big data," Ph.D. dissertation, Dept. Comput. Sci., Udima, Madrid, Spain, 2016.
[24] E. González, "Bases de datos NoSQL: Desarrollo de un prototipo de aplicación basada en MongoDB," Ph.D. dissertation, Dept. Comput. Sci., Udima, Madrid, Spain, 2016.
[25] J. Olivera, "Big data: Aplicación y utilidad en el sistema alimentario," Ph.D. dissertation, Dept. Comput. Sci., Udima, Madrid, Spain, 2016.
[26] F. Sestayo, "Análisis de datos semiestructurados mediante big data y map reduce," Ph.D. dissertation, Dept. Comput. Sci., Udima, Madrid, Spain, 2016.
[27] A. Aguilar, "Bases de datos NoSQL: Desarrollo de un prototipo basado en MongoDB y Apache cassandra," Ph.D. dissertation, Dept. Comput. Sci., Udima, Madrid, Spain, 2019.
[28] F. J. Moreno, "Sistema big data para mejorar los rendimientos agrícolas en castilla y León," Ph.D. dissertation, Dept. Comput. Sci., Udima, Madrid, Spain, 2019.
[29] P. Pyzel, "Ampliación de un sistema de Big data para mejorar los rendimientos agrícolas con objetivo de realizar previsiones de necesidades de agua tratada en países con escasez de recursos hídricos," Ph.D. dissertation, Dept. Comput. Sci., Udima, Madrid, Spain, 2019.
[30] S. M. Bulger and L. D. Housner, "Modified Delphi investigation of exercise science in physical education teacher education," *J. Teach. Phys. Edu.*, vol. 26, no. 1, pp. 57–80, Jan. 2007.
[31] R. D. Penfield and P. R. Giacobbi, Jr., "Applying a score confidence interval to Aiken's item content-relevance index," *Meas. Phys. Edu. Exerc. Sci.*, vol. 8, no. 4, pp. 213–225, Dec. 2004.
[32] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM1.0 Step-by-step data mining guide," SPSS, Chicago, IL, USA, Tech. Rep., 2000.

JUAN A. LARA (Member, IEEE) received the Ph.D. degree in computer science and two post graduate master's degree in information technologies and emerging technologies to develop complex software systems from the Technical University of Madrid, Spain.

He is an Associate Professor and Research Scientist with Madrid Open University, UDIMA, Spain. He is currently the Director of the Group of Research in Knowledge Management and Engineering. He is the author of more than five online education books. He has published some book chapters and papers on several international conferences, and taken part in national and international research projects. He is the author of more than 25 articles published in international impact journals. His research interests are in computer science include data mining, knowledge discovery in databases, data fusion, artificial intelligence, and e-learning.

AUREA ANGUERA DE SOJO received the B.A. degree in law, the B.A. degree in economics from the Universidad Pontificia de Comillas (ICADE), Madrid, Spain, and the Ph.D. degree in computer science from the Universidad de Coruña, Spain.

She has worked at the Universidad Nacional de Educación a Distancia (UNED), since 1996, until now teaching about human research, law, and e-commerce. Since 2002, she has been an Associate Professor with the Computer Systems Department, Universidad Politécnica de Madrid (UPM), imparting classes in e-commerce, social, professional, legal and ethics aspects of engineering, and in business intelligence. Her research interests include ICT social, economic and legal implications, privacy, and artificial intelligence.

SHADI ALJAWARNEH (Member, IEEE) received the B.Sc. degree in computer science from Jordan Yarmouk University, the M.Sc. degree in information technology from Western Sydney University, and the Ph.D. degree in software engineering from Northumbria University, U.K. He is currently a Full Professor of software engineering with the Jordan University of Science and Technology. He has presented at and been on the organizing committees for a high number of international conferences and is a board member of the International Community for the ACM, IEEE, Jordan ACM Chapter, ACS, and others. A good number of his articles have been selected as "Best Papers" at conferences and journals. Also, he has served as the Conference Chair, TPC Chair for a good number of international conferences. Furthermore, he is an Associate Editor of the *Computers and Electrical Engineering Journal* (Elsevier) and a guest editor for many journals special issues.

ROBERT P. SCHUMAKER (Senior Member, IEEE) received the B.Sc. degree in civil engineering from the University of Cincinnati, in 1997, the MBA degree in management and international business from the University of Akron, in 2001, and the Ph.D. degree in management from the University of Arizona, in 2007.

He has authored a book on *Sports Data Mining*, several book chapters, multiple journal articles in DSS, ACM TOIS, CACM, and JASIST. He had his research featured in the Wall Street Journal and numerous other media outlets. He is the George W. and Robert Pirtle Endowed Professor at the Department of Computer Science, University of Texas at Tyler (UT Tyler), and the Director of the Data Analytics Lab, Soules College of Business. His overall research interests involve using data science to solve very large and complex business problems. These interests further branch into data mining, machine learning, natural language processing, sentiment analysis, system building, and textual analytics. In particular, he focuses on the areas of textual/financial prediction, sports analytics, and healthcare informatics. He is a member of the ACM. He is also an Associate Editor of the *Decision Support Systems Journal* (Elsevier) and speaker through the ACM Distinguished Speakers Program (April 2013 – April 2019).

BASSAM AL-SHARGABI (Member, IEEE) received the B.Sc. degree in computer science from the Applied Science University, Jordan, in 2003, and the M.Sc. and Ph.D. degrees in computer information systems from the Arab Academy for Banking and Financial Sciences, Jordan, in 2004 and 2009, respectively.

He is currently an Associate Professor with the Department of Computer Information System, Faculty of Information Technology, Middle East University, Amman, Jordan. His current research interests are in natural language processing, machine learning, the Internet of Things, and service-oriented architecture.