Big Data Education At The Chilean Academy: Is This Possible?

Cristian Vidal-Silva, Erika Madariaga, Claudia Jiménez, Luis Urzúa

Abstract: Technology evolves, and the human being presents a growing need for the use and generation of large volumes of information and data (Big Data). Working with Big Data with traditional computer systems is not feasible: Including new knowledge and technology of Big Data for inclusion in professional computer and computer education is necessary. The main objective of this work is to answer whether or not the Chilean academy is prepared to train specialists in Big Data. In addition to describing theoretical and practical components of Big Data along with introducing an essential tool of the subject, this work defines and presents the results of a survey to explore and analyze the reality of the academy in Chile regarding the degree of viability to train and train professionals competent in Big Data. Necessary conditions for developing Big Data competences in the Chilean academy require more adjustments. Specifically, the Chilean academy needs to adopt Big Data topics and solutions for developing those competencies in future professionals.

Index Terms: Big Data, Technology, Large Volumes of Information, Education, Competences, Chilean Academy.

1. INTRODUCTION

A data is a symbolic representation of some observable property of an object, event, and environment that without a contextualization and purpose is not relevant for decision making [1]. According to [2], the information corresponds to data that adds value for the understanding of a particular topic, of great relevance for organizations in decision-making processes and knowledge generation. Such as Liew [3] argues, knowledge is a body of information to create and increase value for the company. Such as [4] highlight, data processing originates information, and information processing generates knowledge. The pyramids DIKW (Data, Information, Knowledge: Knowledge, Wisdom) [3] and DIKIW (Data, Information, Knowledge, Intelligence, Wisdom) [1] [5] consider wisdom and intelligence as results of knowledge processing, and wisdom as a result of intelligence processing, respectively. The basis of all these contexts is data and its processing to obtain information. According to White [6], databases that use structured SQL query language (Structured Query Language) work on structured information, the knowledge area, and basic training in technical and professional majors in the areas of information science, computer science, and computing and businesses for years in Chile. Information with less structure is common in the last years [7]. Moreover, the volume of information increases over the limits of traditional computer systems. Therefore, education in the areas previously mentioned is really necessary. For example, internet computing systems such as Facebook and web hosting servers work with semi-structured and unstructured data, usually used for the new knowledge and information generation. Today we live in the Big Data society, that is, in a society that requires working with large volumes of data and information. Such as White [6] argues, the amount of data generated by machines, as part of the Internet of Things, becomes even more extensive than the data that humans produce. For example, authentication and access records (logs), radiofrequency readers, network sensors, GPS sensors in vehicles, web transaction logs contribute significantly to the growth of Big Data. The good news is that Big Data is here, but technical skills for storage and knowledge for analysis are not developed in Chile yet. In higher education in Chile and Latin America, except for research papers such as [8-10], there is no clear evidence or the tendency for the development of skills Big Data domain. The training of Engineers in Chile does not yet explicitly include Big Data courses (unless the Big Data topic were like this in so-called optional or specialization workshops). The objective of this work is to present and describe the current trends of processing large volumes of data and information by the use of tools and computational methodologies for that purpose. Then, we detail an exploratory study and results of a survey to know the availability of including Big Data topics into Chilean higher education is or not feasible. Software and hardware of traditional computer systems present space and speed issues for working with the large volume approaches appear as solutions, even though implementing the former represent a high cost, and the latter do not present a great software support for easily implementing computer systems.

In the history of computing, Google represents a search engine characterized by the quality and efficiency of its results. Dean and Ghemawat [11] presented details of the MapReduce approach as the base of the computing platform of Google. Since then, the world of free software works on products that achieve the performance of Google products. Google's MapReduce platform is the basis of the free Hadoop tool. Likewise, Google's Preguel represents the basis of free tools such as Giraph [9-10] [12] and Spark [13-14]. Such as Vidal et al. [10] discuss, MapReduce represents an original Google programming methodology [11] [15] for distributed computing over large volumes of data or Big Data, with a wide dissemination and adoption of its free and open-source implementation Hadoop [6] [16-17]. The divide-and-conquer algorithmic technique is the algorithmic base of MapReduce [6] [18-19]. Thus, MapReduce as a distributed system, divides the data into smaller portions for parallel processing, either in parallel by multiple threads working in a processor, by multiple processors working in a multiprocessor machine, or by multiple working machines in a cluster or network of machines

- *Cristian Vidal-Silva, professor at Departamento de Administración, Facultad de Economía y Negocios, Universidad Católica del Norte, Antofagasta, Chile. E-mail: cristian.vidal@ucn.cl*
- *Erika Madariaga, director of Ingeniería Informática, Facultad de Ingeniería, Ciencia y Tecnología, Universidad Bernardo O'Higgins, Santiago, Chile. E-mail: erika.madariaga@ubo.cl*
- *Claudia Jiménez, director of Ingeniería Civil Informática, Facultad de Ingeniería y Negocios, Universidad Viña del Mar, Viña del Mar, Chile. E-mail: cjimenez@uvm.cl*
- *Luis Urzúa, professor at Escuela de Kinesiología, Facultad de Salud, Universidad Santo Tomás, Talca, Chile. E-mail: lurzua@santotomas.cl*

[6]. The final output of this process is the combination of intermediate results of each of its workers (mappers and reducers). In practice, MapReduce is a distributed computing framework that allows avoiding distributed programming costs, such as guaranteeing the sending and receiving of messages.

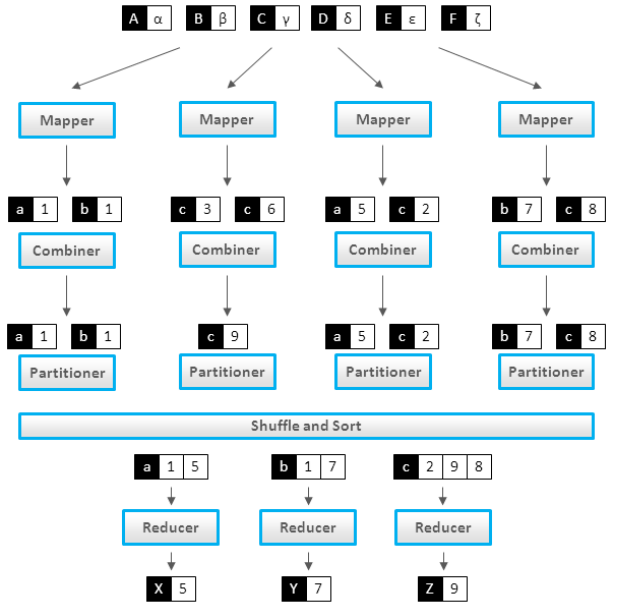

***Fig. 1**. Functioning architecture of MapReduce.*

Figure 1 [10] illustrates the four main operating architecture steps of a MapReduce solution in the Hadoop MapReduce philosophy: i) Fractioning and delivering of data fractions to each of the mappers. ii) Map modules that perform the map functions for identifying relevant elements to send them to the next stage. iii) The organization stage that organizes data and groups intermediate results for distributing them to the reduction stage. iv) Reduce modules where the Reduce functions are executed to compact and summarize the results to write them on disk. Thus, in Hadoop and MapReduce, Mappers objects execute map functions, and Reducer instances execute reduce functions. For more detail of the MapReduce operation and philosophy, we suggest reviewing (White, 2015; Vidal et al., 2018). This paper structures as follows: Section 2 describes methodology of this study. Section 3 presents and discuss the main results of this study. Conclusions finally concludes and presents future work ideas.

2 METHODOLOGY

Table 1 details a survey applied to teachers belonging to the area of technologies and information technology of some institutions of higher education in Chile. Correctly, this survey was applied to 15 teachers belonging to different Chilean higher education entities. The survey was conducted at the end of 2016 and the beginning of 2017 and is composed of 15 questions classified in 4 segments. This study was carried out through the Google Docs platform under the name of the Current Status of Big Data in the Chilean Academy [20]. Figure 2 shows part of the tool interface for identification purpose. The study Current Status of Big Data in the Chilean Academy, as its name implies, seeks to measure and know the current reality of Big Data in Chilean higher education, as shown in Table 1, by quantifying the following variables: previous use of computer and Big Data technology, level of specialization in computer and Big Data technologies, level of relevance of Big Data for your academic institution. Tables 2, 3, and 4 present the questions for the variables of this study along with their answers using a Likert scale [21] to measure the feasibility for training in Big Data in the Chilean academy.

TABLA I

SURVEY FOR TEACHERS IN THE AREA OF TECHNOLOGY AND INFORMATION TECHNOLOGY OF HIGHER EDUCATION ENTITIES IN CHILE.

#	*Focus*	*Questions*	*Answer Type*
1	Personal Data Surveyed.	1 – 5	Single option
2	The use of skills, computer technologies, and Big Data.	6 – 8	Multiple options
3	The specialization level of Information Technology and Big Data.	9 – 12	Single option
4	The level of relevance of Big Data	13 – 15	Single option

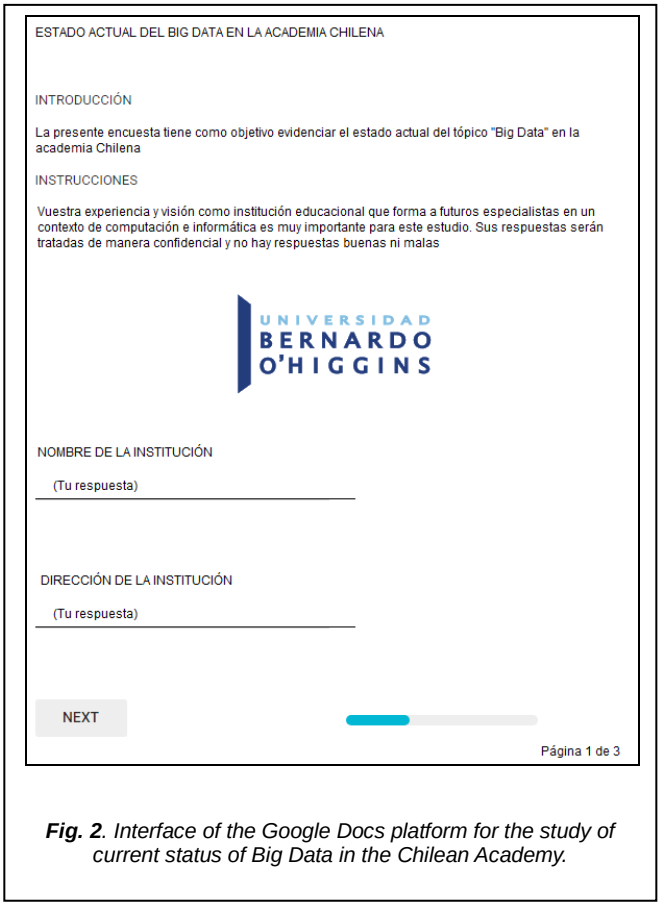

***Fig. 2**. Interface of the Google Docs platform for the study of current status of Big Data in the Chilean Academy.*

3 RESULTS

According to the results of the survey, 93% of respondents represent university teachers, and the remaining 7% belong to another type of higher education institution, as shown in Figure 3. In this context, as seen in Figure 4, 40% of respondents fulfill the role of only a teacher, who do not devote themselves to research, a semi-limitation towards the acquisition and development of new themes. Likewise, 40% of the teachers surveyed fulfill the role of teacher and researcher who are prepared to investigate new technologies, techniques, or methodologies such as Big Data in theory. 20% of respondents indicate that their role refers to another, the one that could only be a coordinator, director, or another similar role. 0% of respondents fulfill a research-only role. When asked about the tools or techniques used by the institution regarding the processing of data and information (1st question of table 3), in the field totally oriented to Big Data, 33% of respondents replied that the most used were Bases of NoSQL Data, even if the adoption of a philosophy of how MapReduce is key to understanding contexts, types of problems and possible results to solve by working with large volumes of data. Besides, 73% of respondents responded that the most used tool not related to Big Data corresponds to SQL databases, although that tool mainly allows working with volumes of structured and relational information no of the size of Big Data itself.

TABLA II
QUESTIONS ABOUT THE USE OF SKILLS, COMPUTER TECHNOLOGIES, AND BIG DATA.

Question	Answer Type	Answer Options
In the following table, select those techniques, technologies, or tools that your institution has ever used.	Inclusive Options	Database (BD) NoSQL - SQL Database – Hadoop MapReduce - R - SAS - Machine Learning - Java - Hive QL - Python - Mahout - Tableu - Julia - Ipython - Ruby - Qlikview - Other (indicate)
In the following table, select those statistical skills that your institution has ever used.	Inclusive Options	Methodology for Big Data processing - Standards for Big Data processing - Use of statistical software such as Excel, SAS, SPASS or similar - Data management skills including documentation, registration, and access control - Ability to work with text analysis - Data Mining - Other (indicate)
In the following table, select those other skills developed by your institution.	Inclusive Options	Communication - Creativity for problem-solving - Teamwork - Initiative - Privacy - Data governance - Ethics - Other (indicate)

***Figures 5 and 6** detail the academic productivity in the Big Data topic at the institutions of higher education of the respondents. Academic productivity considers publications in indexed or non-indexed journals, presentation of papers in academic congresses, or some other type of event on the subject of Big Data.*

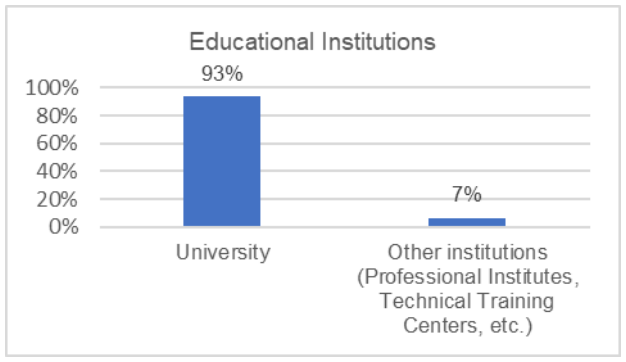

***Fig. 3**. Higher education institutions of respondents.*

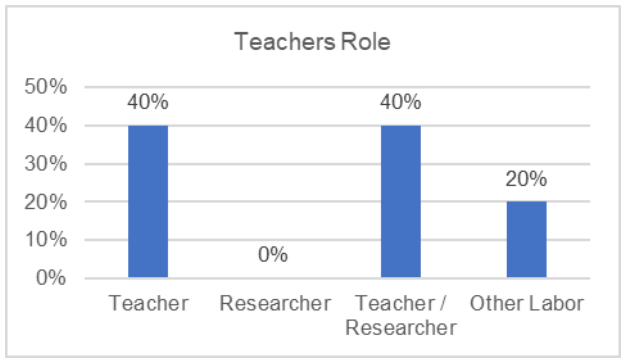

***Fig. 4**. Main role of respondents.*

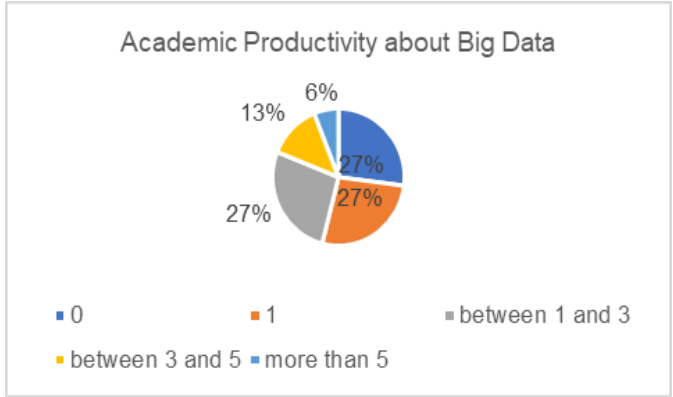

***Fig. 5.** Academic productivity of the institutions of the interviewees in the Big Data topic.*

Academic productivity corresponds to the results of the last questions in table 4. These results show that the academic productivity of the respondents in the Big Data topic is very low (practically zero). There is a clear correlation between the role of the respondents and their academic productivity. Thus, people who are non-exclusive researchers experienced little updating in new technologies such as Big Data. These results are consistent with what Figure 7 shows, where only 13% of the educational institutions of the respondents consider the Big Data topic to be highly relevant.

TABLA IV
QUESTIONS ABOUT THE USE OF SKILLS, COMPUTER TECHNOLOGIES, AND BIG DATA.

Question	Answer Options
How important are the techniques and tools of BIG DATA in your career plans and programs? (Select your alternative)	Nothing - Low - Medium - Advanced - Expert
Number of application or research work about BIG DATA that was carried out in your institution (check according to quantity)	0 - 1 - between 1 and 3 - between 3 and 5 - more than 5
Number of academic events in which the BIG DATA topic has been present, in which you have participated (check according to quantity)	0 - 1 - between 1 and 3 - between 3 and 5 - more than 5

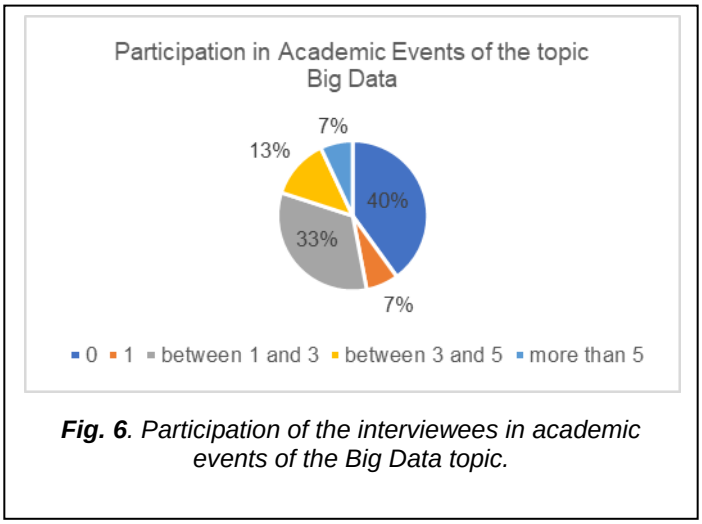

***Fig. 6**. Participation of the interviewees in academic events of the Big Data topic.*

Clearly, due to the massive use of technology and the presence of large volumes of data (Big Data), competencies related to Big Data are essential in various current areas such as business and education [22-23]. Such as Dede [23] mentions, education research can benefit greatly from Big Data and the current computer revolution. Now the availability of online resources related to educational technology expands research opportunities in learning, including gender, ethnicity, economic situation, among other variables. Education and research in Chile should consider these resources for the teaching and learning process in the development of Big Data skills.

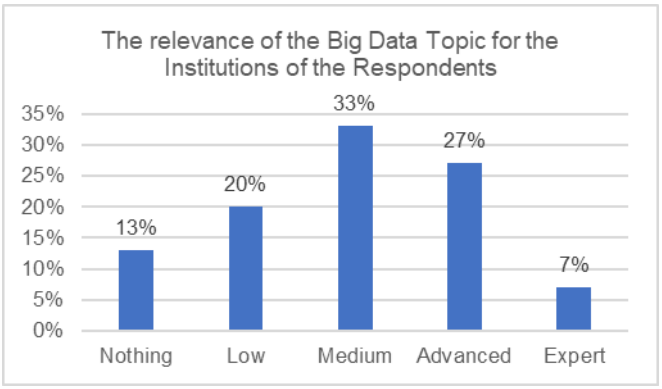

Fig. 7. *Participation of the interviewees in academic events of the Big Data topic.*

According to the work of Williams [24], English analytical learning and the massive use of information or Big Data play a preponderantly positive role in the development of competencies in university students. As previously noted, this invites universities in Chile to use guided learning platforms as tools to support Big Data thematic courses (and also any other competition). The incorporation of Big Data in the academy requires the inclusion of the topics of analysis, design, implementation, and testing of solutions for the software engineering and computer areas as well as for any other area that works with information which is current for any area of knowledge. This work presented the results of a study about the reality of Big Data in the Chilean academy. According to these results, in the academy in Chile, even when there are the bases to enter the world of Big Data, a structural change in theoretical and practical content is required to achieve competent professional training in Big Data. Besides, the effective realization of such change requires a definition of skills and learning outcomes, as well as levels of development. Including Big Data in the computer academy in Chile is being similar to the incorporation of object-oriented software development. For example, one of the authors did not develop object-oriented competencies in the undergraduate (1998-2003), and this paradigm existed in the world industry and academy since the late 1960s [25] and mid- 80, respectively. Big Data is known and popular for Google since 2004 [11], and, with the elapsed time, this work shows the reality of the academy in Chile in the viability of training competent professionals in Big Data.

TABLA III
QUESTIONS ABOUT THE DEGREE OF SPECIALIZATION IN COMPUTER TECHNOLOGY AND BIG DATA.

Question	Answer Options	Items
To indicate the degree of specialization that you reach in working with the following computational tools. If you have not used any of them, the degree of knowledge "Nothing" seems an adequate option.	Nothing - Low - Medium - Advanced - Expert	NoSQL Database - SQL Database - Hadoop - MapReduce - R - SAS - Machine Learning - Java - Hive QL - Python - Mahout - Tableu - Julia - Ipython - Ruby - Qlikview - Other (indicate)
To indicate the degree of specialization that you reach in working with the following statistical skills. If you do not possess any specialization skill, the degree of knowledge "Nothing" seems an adequate option.	Nothing - Low - Medium - Advanced – Expert	Methodology for Big Data processing - Standards for Big Data processing - Use of statistical software such as Excel, SAS, SPSS or similar - Data management skills including documentation, registration, and access control - Ability to work with analysis of data - Data mining - Other (indicate)
To indicate the degree of specialization that you reach working with the following skills. If you do not possess any of these skills, the degree of knowledge "Nothing" seems an adequate option.	Nothing - Low - Medium - Advanced – Expert	Communication - Creativity for problem-solving - Teamwork - Initiative - Privacy - Data governance - Ethics - Other (indicate)

4 FINAL DISCUSSION

The incorporation of Big Data in the academy requires the inclusion of the topics of analysis, design, implementation, and testing of solutions for the software engineering and computer areas as well as for any other area that works with information which is current for any area of knowledge. This work presented the results of a study about the reality of Big Data in the Chilean academy. According to these results, in the academy in Chile, even when there are the bases to enter the world of Big Data, a structural change in theoretical and practical content is required to achieve competent professional training in Big Data. Besides, the effective realization of such change requires a definition of skills and learning outcomes, as well as levels of development. Including Big Data in the computer academy in Chile is being similar to the incorporation of object-oriented software development. For example, one of the authors did not develop object-oriented competencies in the undergraduate (1998-2003), and this paradigm existed in the world industry and academy since the late 1960s [25] and mid- 80, respectively. Big Data is known and popular for Google since 2004 [11], and, with the elapsed time, this work shows the reality of the academy in Chile in the viability of training competent professionals in Big Data.

5 CONCLUSIONS

This work defined and presented the and analysis and results of a study about the feasibility of training competent

professionals in the areas of computer science and software engineering in the subject of Big Data in the academy in Chile. Then, this work argued that there are not yet all the conditions for the training of IT professionals specialized in Big Data, despite the massive relevance and dissemination of the subject. This work suggested that the inclusion of Big Data topics is necessary for the training of new generations of engineers in the areas of computer science and software engineering in Chile. Currently, the demand for professionals specialized in Big Data is growing, and the mastery of these topics is necessary for the development of solutions to current Big Data problems. The programming, modeling, and software engineering bases already exist for the successful adoption of Big Data in the Chilean academy. Hence, it is required to coordinate its inclusion for the training of professionals and researchers of Big Data topics.

REFERENCES

[1] S. Baskarada, and A. Koronios, “Data, Information, Knowledge, Wisdom (DIKW): A Semiotic Theoretical and Empirical Exploration of the Hierarchy and its Quality Dimension”, Australasian Journal of Information Systems, vol. 18, no. 1, Nov 2013, doi: 10.3127/ajis.v18i1.748.

[2] D. Chaffey, and S. Wood, “Business Information Management: Improving Performance Using Information Systems”, Harlow, FT Prentice Hall, pp. 102-110, 2005.

[3] A. Liew, “Understanding Data, Information, Knowledge and their interrelationships”, Journal of Knowledge Management Practice, vol. 7, June 2007.

[4] T. Davenport, and L. Prusak, “Working Knowledge: How Organizations Manage What They Know”, Journal Ubiquity, ACM, USA, Aug 2000, doi: 10.1145/347634.348775.

[5] A. Liew, “DIKIW: Data, Information, Knowledge, Intelligence, Wisdom and their Interrelationships”, Journal Business Management Dynamics, vol. 10, no. 2, pp. 49-62, Apr 2013.

[6] T. White, “Hadoop: The Definitive Guide”, O’Reilly, 4th Ed., Sebastopol, CA, USA, 2015, ISBN: 1491901632.

[7] A. Gandomi, and M. Haider, “Beyond the hype”, International Journal of Information Management, vol. 35, no. 2, pp. 137–144, Apr 2015, doi: 10.1016/j.ijinfomgt.2014.10.007.

[8] J. J. Camargo, J. F. Camargo, and L. Joyanes, “Conociendo Big Data”, Revista Facultad de Ingeniería, vol. 24, no. 38, pp. 63-77, Colombia, Apr 2015.

[9] S. Valenzuela, C. Vidal, J. Morales, and L. López, “Ejemplos de Aplicabilidad de Giraph y Hadoop para el Procesamiento de Grandes Grafos”, Información Tecnológica, vol. 27, no. 5, pp. 171-180, Chile, Sept/Oct 2016, doi: 10.4067/S0718-07642016000500019.

[10] C. Vidal, M. Bustamante, M. Lappo, and M. Núñez, “En la Búsqueda de Soluciones MapReduce Modulares para el Trabajo con BigData: Hadoop Orientado a Aspectos”, Información Tecnológica, vol. 29, no. 2, pp. 133-140, Chile, Mar/Apr 2018, doi: 10.4067/S0718-07642018000200133.

[11] J. Dean, and S. Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters”, in Proceedings of 6th Conference on Symposium on Operating Systems Design & Implementation OSDI 2004, San Francisco, CA, USA, pp. 137-150, 2004.

[12] R. Shaposhnik, C. Martella, and D. Logothetis, “Practical Graph Analytics with Apache Giraph”, Apress, USA, 2015, ISBN: 1484212525.f

[13] M. Guller, “Big Data Analytics with Spark”, Apress – Spring, New York, NY, USA, 2015, ISBN: 978-1-4842-0965-3.

[14] J. Shi, Y. Qiu, U. F. Minhas, L. Jiao, C. Wang, B. Reinwald, and F. Öscan, “Clash of the Titan: MapReduce vs. Spark for Large Scale Data Analytics”, in Procedings of Very Large Data Bases (VLDB), Sept 5th-09th, New Delhi, India, 2016, doi: 10.14778/2831360.2831365.

[15] J. Dean, and S. Ghemawat, “MapReduce: a Flexible Data Processing Tool”, Communications of the ACM, vol. 53, no. 1, pp. 72-77, 2010, doi: 10.1145/1629175.1629198.

[16] Apache, “Apacahe Hadoop”, Apache Hadoop, https://hadoop.apache.org/. 2019.

[17] J. Lin, and C. Dyer, “Data-Intensive Text Processing with MapReduce”, Morgan and Claypool Publishers, USA, 2010, ISBN: 1608453421.

[18] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, “Introduction to Algorithms”, 3rd Edition, The MIT Press, 2009.

[19] A. Hernández, and A. Hernández, “Acerca de la aplicación de MapReduce + Hadoop en el tratamiento de Big Data”, Revista Cubana de Ciencias Informáticas, vol. 9, no. 3, pp. 49-62, July/Sept 2015.

[20] EstudioBigData, “Estado Actual del Big Data en la Academia Chilena”, Universidad Bernardo O’Higgins. https://docs.google.com/forms/d/e/1FAIpQLSczj8C2XV0VrhrazICB2OSc2Cz7JIIebHzQZM-pV0yl29B0BQ/viewform. 2018.

[21] R. Likert, “A Technique for the Measurement of Attitudes”, Archives of Psychology, vol. 140, pp. 1-55, 1932.

[22] M. Huda, A. Maseleno, P. Atmotiyoso, M. Siregar, R. Ahmad, K. Azmi Jasmi, and N. Hisyam Nor Muhamad, “Big Data Emerging Technology: Insights into Innovative Environment for Online Learning Resources”, International Journal of Emerging Technologies in Learning (iJET), vol. 13, no. 01, Jan 2018, doi: 10.3991/ijet.v13i01.6990.

[23] C. Dede, “Next steps for ‘Big Data’ in Education: Utilizing data-intensive research”, Educational Technology Harvard, LVI, vol. 2, pp. 37 – 42, 2016.

[24] P. Williams, “Assessing collaborative learning: big data, analytics and university futures”, Assessment & Evaluation in Higher Education, July 2016, doi: 10.1080/02602938.2016.1216084.

[25] O. J. Dahl, B. Myhrhaug, and K. Nygaard, “Some features of the SIMULA 67 language”, Proceedings of the Second Conference on Applications of Simulations, 29 – 31 Dec 1968.