Article **Journal of Technology and Education**
December 2021, Vol.5 No.14 1-5

Big Data processes and tools learning environment for bachelor's degree students in computer engineering

Entorno de aprendizaje de procesos y herramientas de Big Data para alumnos de licenciaturas en ingeniería en computación

HERNÁNDEZ-CABRERA, Jesús†*, MENDOZA-GONZÁLEZ, Omar and SÁNCHEZ-HERNÁNDEZ Miguel Ángel

Universidad Nacional Autónoma de México, Facultad de Estudios Superiores Aragón, Mexico.

ID 1st Author: *Jesús, Hernández-Cabrera* / **ORC ID:** 0000-0002-7850-858X, **CVU CONACYT ID:** 104580

ID 1st Co-author: *Omar, Mendoza-González* / **ORC ID:** 0000-0002-3492-4549, **CVU CONACYT ID:** 972783

ID 2nd Co-author: *Miguel Ángel, Sánchez-Hernández* / **ORC ID:** 0000-0002-6265-1760, **CVU CONACYT ID:** 326529

DOI: 10.35429/JTAE.2021.14.5.1.5 Received: July 10, 2021; Accepted: December 30, 2021

Abstract

The set of technologies related to the Big Data environment is broad and continues to grow; undergraduate students interested in venturing into learning this type of technology are, in the first instance, overwhelmed because the learning curve is extensive; The objective of the project is to reduce this curve; For which a massive data processing environment has been designed and generated, accompanied by notes with fundamental concepts of Big Data, workshops and theoretical-practical courses, all supervised by academics from FES ARAGÓN. This document shares the experience gained in one year of its application and the various products that are detailed in the extensive. The methodology to meet the objective consisted of the following phases: Documentary research, content design and development, creation and configuration of the environment for Big Data tools, environmental operation tests, generation of teaching material, teaching of courses and workshops. and publication of the materials in a MOOC. The contribution proposed in this article is to socialize the experience obtained from the teaching support and innovation project, in order for other academic centers to take it up, replicate or improve it.

Bigdata, Workshop, Education

Resumen

El conjunto de tecnologías relacionadas al entorno Big Data es amplio y sigue en crecimiento; los alumnos de licenciatura interesados en incursionar en el aprendizaje de este tipo de tecnologías, de primera instancia se ven rebasados debido a que la curva de aprendizaje es extensa; el objetivo del proyecto es reducir esta curva; para lo cual se ha diseñado y generado un entorno de procesamiento de datos masivos, acompañado por notas con conceptos fundamentales de Big Data, talleres y cursos teóricos-prácticos, todo supervisado por académicos de la FES ARAGÓN. En este documento se comparte la experiencia lograda en un año de su aplicación y los diversos productos que se detallan en el extenso. La metodología para cumplir con el objetivo constó de las siguientes fases: Investigación documental, diseño y desarrollo del contenido, creación y configuración del entorno para las herramientas de Big Data, pruebas de operación del entorno, generación del material didáctico, impartición de cursos y talleres y publicación de los materiales en un MOOC. La contribución propuesta en el presente artículo es socializar la experiencia obtenida del proyecto de apoyo e innovación a la docencia, con el fin que otros centros académicos lo retomen, repliquen o mejoren.

Big Data, Taller, Educación

Citation: HERNÁNDEZ-CABRERA, Jesús, MENDOZA-GONZÁLEZ, Omar and SÁNCHEZ-HERNÁNDEZ Miguel Ángel. Big Data processes and tools learning environment for bachelor's degree students in computer engineering. Journal of Technology and Education. 2021. 5-14:1-5.

* Correspondence of the Author (Email: jesushc@unam.mx)
† Researcher contributing as first author.

Introduction

The growing demand for professionals in the area of computing with knowledge of tools, processes and environments related to the processing of large volumes of data has already exceeded the offer provided by higher education institutions.

Therefore, there is a need for early and continuous training for students currently studying engineering in the area of computing, supported by a teaching-learning environment reinforced with the use of tools and theoretical-practical examples directly linked to the technologies required by the labour and professional market.

The project proposes the following hypothesis: a student interested in learning about Big Data will benefit from the attenuation of the learning curve by having access to a limited and formal learning environment with tools and theoretical-practical examples on the particular subject.

In order to test the above hypothesis, a study was carried out on the use of 70 students, in one edition of the workshop and two editions of the course, making use of the configured environment and accessing the support material generated expressly for the project.

Objectives

To generate a teaching-learning environment for Big Data processes, tools and technologies, supported by concrete deliverables such as workshops, courses, didactic material, online resources and the implementation of a distributable environment with Hadoop tools, available to students and academics at the UNAM undergraduate level.

Contribution

Through the project's deliverables, 70 students have been trained in a professional manner with a real approach to the architecture, tools and processes of massive data management, since in the workshops, practices were carried out using a pre-configured environment with the elements of the Hadoop ecosystem, HDFS, MapReduce, YARN, Hive, Sqoop and Flume. The resulting products can be used in research and teaching projects, both at school and professional level.

Multidisciplinary collaboration has been promoted as the new forms of data and knowledge generation require a distribution of work based on collaborative platforms that demand new and better skills for the resolution of social, industrial, academic and scientific problems at national level.

Development

The following products were developed:

1. Integration of didactic material.

Bibliographic research was carried out covering the topics of interest for the treatment of massive data, from which the following products emerged:

1.1 A tutorial type manual for the course on Big Data concepts and related topics.

1.2 A participant's manual for the introductory course on Big Data.

1.3 A practice workbook which consists of 6 practices which cover:

1.3.1 HDFS, 1.3.2 MapReduce, 1.3.3 Apache Hive, 1.3.4 Apache Sqoop, 1.3.5 Apache Flume, 1.3.6 HUE environment.

2. Building a virtual machine image with the Hadoop environment.

A virtual machine image was built, based on the Ubuntu 18.04 operating system and the Hadoop environment version 3.1.3 and the data ingestion and processing tools used for the courses and workshops were configured, comprising the following: for data ingestion Apache Sqoop 1.4.7, Apache Flume 1.9, for data processing: Apache Hive 3.1.2 and for administration and operation of the environment HUE 4.1 was installed, in addition, as part of the environment a MySQL 5.7.33 server was installed. The configuration can be presented as shown in Figure 1.

3. Practical Workshop.

An edition of the Big Data workshop was given with the following agenda and practical exercises:

ISSN 2523-0360

1. Introduction to Big Data, 2. Apache Hadoop. 3. Hadoop Ecosystem, 4. HDFS, 5. MapReduce, 6. Hive, 7. Flume, 8. Sqoop, 9. Installation of Hadoop, 11. Installation of a DBMS (MySQL), 12. Installation of the Hadoop ecosystem tools (Sqoop, Flume, Hive), 13.

4. Delivery of inter-semester courses

Two editions of the course were given in which the following topics were covered in a theoretical and practical way:

1. Introduction to Big Data, 2.Apache Hadoop
3. HDFS, 5. Mapreduce, 6. Hive, 7. Flume, 8.

5. Building a MOOC resource.

Currently we have a 90% functional version of a MOOC (Massive Online Open Courses) resource that in its first stage is aimed at the UNAM community and in the second stage will be open to the public, users will be able to access all the course material along with the practices and at the end the knowledge will be reinforced with a self-assessment questionnaire.

Description of the methodology

Generate complements or adaptations to digital environments for teaching and learning processes under the following guidelines:

a) Search for information in databases and refereed scientific articles, as well as in the official sites of manufacturers and organisations that manage the development of tools for big data in order to update the content of the theoretical framework of each of the courses and workshops designed, as well as the applications of Big Data concepts in general.

b) Development of the content of the courses, workshops and of the topics that will integrate the didactic material and online resources.

c) Creation and configuration of the working environment for Big Data, consisting of a virtual machine instance containing the configuration of the framework and the tools to be used.

d) Operational testing of the Big Data environment.

e) Generation of the didactic material, which will be included in the repository of digital resources.

f) Delivery of the courses and workshops, both face-to-face and distance learning.

h) Publication through a MOOC of the didactic materials of the Big Data course and workshop.

i) Generation of a final report.

Graphics

Figure 1 represents the components of the ecosystem installed in the virtual machine, on which both the practical workshop and the inter-semester courses were based.

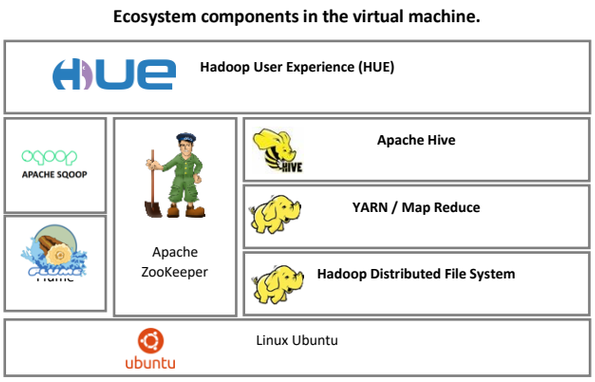
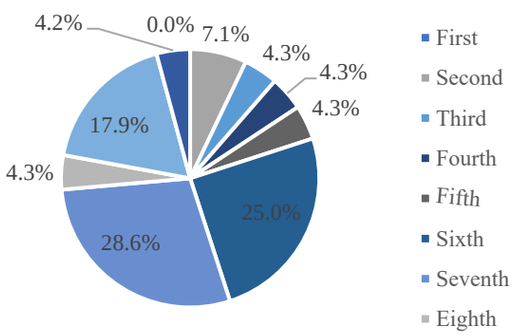
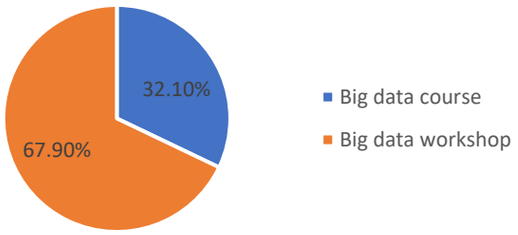


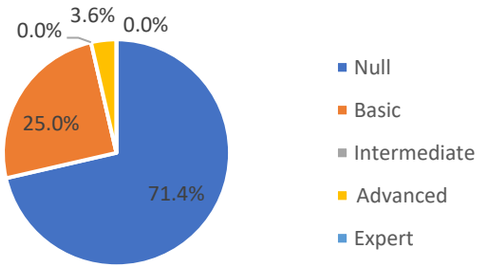
Figure 1 Ecosystem components in the virtual machine
Source: Own elaboration



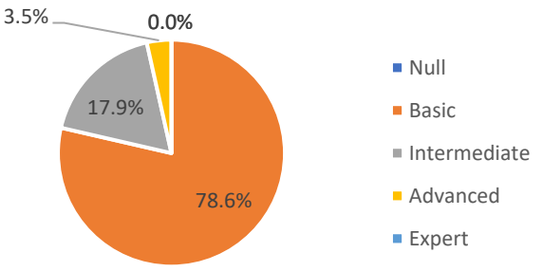
Graphic 1 Percentage of responses to the question "Which semester were you in when you took the course?"
Source: Own elaboration



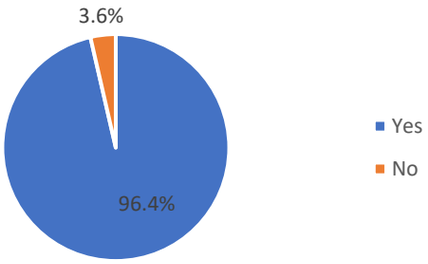
Graphic 2 Percentage of responses to the question "Which course did you take?"
Source: Own elaboration



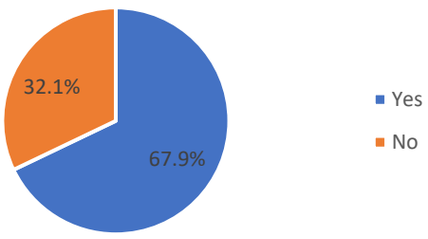
Graphic 3 Percentage of responses to the question "What was your previous level of knowledge of big data?"
Source: Own elaboration



Graphic 4 Percentage of responses to the question "After the course/workshop, how do you consider your level of knowledge of the Hadoop ecosystem?"
Source: Own elaboration



Graphic 5 Percentage of responses to the question "Do you consider continuing to learn Big Data topics?"
Source: Own elaboration



Graphic 6 Percentage of responses to the question "Did the course serve as a basis for implementing your Big Data knowledge in other projects?"
Source: Own elaboration

Results

The population was divided into two important parts, the students enrolled in the Big Data Course and those enrolled in the Big Data Workshop. The differences between the second and the first one lie in the fact that the workshop was held in person and focused on the configuration of the Hadoop modules and tools, with a population mainly of assistants from the intermediate and final semesters of the ICO degree course, On the other hand, the first one was given online and taking as a reference the Big Data Workshop, it was requested that only students from intermediate semesters onwards enrolled in the course, given that students from the third semester onwards do not have basic knowledge of Linux and Database, but they do have a basic knowledge of OOP. A diagnostic instrument was designed where it is observed that the two populations indicate that they have obtained basic knowledge of the Hadoop ecosystem, data processing and ingestion, and a response of over 95% of students who wish to continue learning Big Data and 70% of them managed to implement their knowledge in other projects. It is important to highlight that six of the students who took the workshop also participated as support for the project instructors and were responsible for guiding and advising the students in the courses, which is why they reinforced the knowledge acquired in the workshop.

Conclusions

With this project, a Big Data learning community has been started at the Faculty of Higher Education Aragón. The development of material, a practice environment and course material was achieved.

The practical reinforcement helps students to increase their level of learning and their interest in continuing to deepen their knowledge of the subject.

Based on the foundations of this project, a second stage is being worked on, which consists of the creation and configuration of a cluster using simple and low-cost board equipment that supports massive data processing, so that students can potentially execute Big Data projects related to problems such as urban impact, pothole problems, municipal traffic, industrial production lines, monitoring of social networks, among others.

ISSN 2523-0360

In this second stage, it will be possible to add new tools for ingesting, analysing and visualising massive data in order to expand the academic resources available to students.

References

Chang, Fay & Dean, Jeffrey & Ghemawat, Sanjay & Hsieh, Wilson & Wallach, Deborah & Burrows, Michael & Chandra, Tushar & Fikes, Andrew & Gruber, Robert. (2008). Bigtable: A Distributed Storage System for Structured Data. ACM Trans. Comput. Syst.. 26. 10.1145/1365815.1365816.

Davoudian, A., & Liu, M. (2020). Big Data Systems. ACM Computing Surveys, 53(5),1–39.

Dean, Jeffrey & Ghemawat, Sanjay. (2004). MapReduce: Simplified Data Processing on Large Clusters. Communications of the ACM. 51. 137-150. 10.1145/1327452.1327492.

Hive Documentación. [Online]. Available: https://cwiki.apache.org/confluence/display/Hive/LanguageManual

J. Moreno, E. B. Fernandez, M. A. Serrano and E. Fernández-Medina, "Secure Development of Big Data Ecosystems," in IEEE Access, vol. 7, pp. 96604-96619, 2019, doi: 10.1109/ACCESS.2019.2929330.

S. Ghemawat, H. Gobioff, S. Leung. "The Google file system, " In Proc. of ACM Symposium on Operating Systems Principles, Lake George NY, Oct 2003, pp 29-43.

Sam R. (2017) Expert Hadoop Administration, Managing Turing and Securing SPARK,YARN and HDFS. United States of America:Addison-Wesley Hadoop Documentación. [Online]. Available: https://hadoop.apache.org/docs/r3.1.4/

Sanagustin, M. M. (2016). Mooc Marker. Online, available at: http://www.moocmaker.org/wp-content/files/D1.1-InformeMOOCLatam-vFINALDEFINITIVO_Spanish.pdf

Saroha, M., & Sharma, A. (2019). Big Data and Hadoop Ecosystem: A Review. *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*.

Ting, K., & Cecho, J. (2013). Apache Sqoop Cookbook. O'Reilly Media.

White, T. (2015). Hadoop: The Definitive Guide. O'Reilly & Associates.

ISSN 2523-0360