San Pablo Catholic University (UCSP)
Undergraduate Program in
Computer Science
SILABO

Universidad Católica San Pablo
2021-II

CS370. Big Data (Mandatory)

1. General information

1.1 School	:	Ciencia de la Computación
1.2 Course	:	CS370. Big Data
1.3 Semester	:	9^{no} Semestre.
1.4 Prerrequisites	:	• CS272. Databases II. (5^{th} Sem) • CS3P1. Parallel and Distributed Computing . (8^{th} Sem)
1.5 Type of course	:	Mandatory
1.6 Learning modality	:	Virtual
1.7 Horas	:	1 HT; 2 HP; 2 HL;
1.8 Credits	:	3

2. Professors

Lecturer

- Alvaro Henry Mamani-Aliaga <ahmamani@ucsp.edu.pe>
  - PhD in Ciencia de la Computación, UNSA, Perú, 2019.
  - MSc in Ciencia de la Computación, IME-USP, Brasil, 2011.

3. Course foundation

Nowadays, knowing scalable approaches to processing and storing large volumes of information (terabytes, petabytes and even exabytes) is fundamental in computer science courses. Every day, every hour, every minute generates a large amount of information which needs to be processed, stored, analyzed.

4. Summary

1. Introducción a Big Data 2. Hadoop 3. Procesamiento de Grafos en larga escala

5. Generales Goals

- That the student is able to create parallel applications to process large volumes of information
- That the student is able to compare the alternatives for the processing of big data
- That the student is able to propose architectures for a scalable application

6. Contribution to Outcomes

This discipline contributes to the achievement of the following outcomes:

a) An ability to apply knowledge of mathematics, science. (**Usage**)

b) An ability to design and conduct experiments, as well as to analyze and interpret data. (**Usage**)

i) An ability to use the techniques, skills, and modern computing tools necessary for computing practice. (**Usage**)

j) Apply the mathematical basis, principles of algorithms and the theory of Computer Science in the modeling and design of computational systems in such a way as to demonstrate understanding of the equilibrium points involved in the chosen option. (**Usage**)

7. Content

UNIT 1: Introducción a Big Data (15)

Competences: a,b,i

Content	**Generales Goals**
• Overview on Cloud Computing • Distributed File System Overview • Overview of the MapReduce programming model	• Explain the concept of Cloud Computing from the point of view of Big Data[Familiarity] • Explain the concept of Distributed File System [Familiarity] • Explain the concept of the MapReduce programming model[Familiarity]

Readings: Coulouris et al. (2011)

UNIT 2: Hadoop (15)

Competences: a,b,i

Content	**Generales Goals**
• Hadoop overview. • History. • Hadoop Structure. • HDFS, Hadoop Distributed File System. • Programming Model MapReduce	• Understand and explain the Hadoop suite [Familiarity] • Implement solutions using the MapReduce programming model. [Usage] • Understand how data is saved in the HDFS. [Familiarity]

Readings: Hwang, Dongarra, and Fox (2011), Buyya, Vecchiola, and Selvi (2013)

UNIT 3: Procesamiento de Grafos en larga escala (10)	
Competences: a,b,i	
Content	**Generales Goals**
• Pregel: A System for Large-scale Graph Processing. • Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud. • Apache Giraph is an iterative graph processing system built for high scalability.	• Understand and explain the architecture of the Pregel project. [Familiarity] • Understand the GraphLab project architecture. [Familiarity] • Understand the architecture of the Giraph project. [Familiarity] • Implement solutions using Pregel, GraphLab or Giraph. [Usage]
Readings: Low et al. (2012), Malewicz et al. (2010), Baluja et al. (2008)	

8. Methodology

El profesor del curso presentará clases teóricas de los temas señalados en el programa propiciando la intervención de los alumnos.

El profesor del curso presentará demostraciones para fundamentar clases teóricas.

El profesor y los alumnos realizarán prácticas

Los alumnos deberán asistir a clase habiendo leído lo que el profesor va a presentar. De esta manera se facilitará la comprensión y los estudiantes estarán en mejores condiciones de hacer consultas en clase.

9. Assessment

Continuous Assessment 1 : 20 %

Partial Exam : 30 %

Continuous Assessment 2 : 20 %

Final exam : 30 %

References

Baluja, Shumeet et al. (2008). "Video Suggestion and Discovery for Youtube: Taking Random Walks Through the View Graph". In: *Proceedings of the 17th International Conference on World Wide Web*. WWW '08. ACM: Beijing, China, pp. 895–904. ISBN: 978-1-60558-085-2. DOI: 10.1145/1367497.1367618.

Buyya, Rajkumar, Christian Vecchiola, and S. Thamarai Selvi (2013). *Mastering Cloud Computing: Foundations and Applications Programming*. 1st. Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA. ISBN: 9780124095397, 9780124114548.

Coulouris, George et al. (2011). *Distributed Systems: Concepts and Design*. 5th. Addison-Wesley Publishing Company: USA. ISBN: 0132143011, 9780132143011.

Hwang, Kai, Jack Dongarra, and Geoffrey C. Fox (2011). *Distributed and Cloud Computing: From Parallel Processing to the Internet of Things*. 1st. Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA. ISBN: 0123858801, 9780123858801.

Low, Yucheng et al. (Apr. 2012). "Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud". In: *Proc. VLDB Endow.* 5(8), pp. 716–727. ISSN: 2150-8097. DOI: 10.14778/2212351.2212354.

Malewicz, Grzegorz et al. (2010). "Pregel: A System for Large-scale Graph Processing". In: *ACM SIGMOD Record*. SIGMOD '10, pp. 135–146. DOI: 10.1145/1807167.1807184.