

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



**CS370. Big Data (Mandatory)**

**1. General information**

1.1 School	:	Ciencia de la Computación
1.2 Course	:	CS370. Big Data
1.3 Semester	:	9 <sup>no</sup> Semestre.
1.4 Prerequisites	:	<ul style="list-style-type: none"><li>• CS272. Databases II. (5<sup>th</sup> Sem)</li><li>• CS3P1. Parallel and Distributed Computing . (8<sup>th</sup> Sem)</li></ul>
1.5 Type of course	:	Mandatory
1.6 Learning modality	:	Face to face
1.7 Horas	:	1 HT; 2 HP; 2 HL;
1.8 Credits	:	3

**2. Professors**

**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**

<b>UNIT 1: Introducción a Big Data (15)</b>	
<b>Competences: a,b,i</b>	
<b>Content</b>	<b>Generales Goals</b>
<ul style="list-style-type: none"> <li>• Overview on Cloud Computing</li> <li>• Distributed File System Overview</li> <li>• Overview of the MapReduce programming model</li> </ul>	<ul style="list-style-type: none"> <li>• Explain the concept of Cloud Computing from the point of view of Big Data[Familiarity]</li> <li>• Explain the concept of Distributed File System [Familiarity]</li> <li>• Explain the concept of the MapReduce programming model[Familiarity]</li> </ul>
<b>Readings:</b> Coulouris et al. (2011)	

<b>UNIT 2: Hadoop (15)</b>	
<b>Competences: a,b,i</b>	
<b>Content</b>	<b>Generales Goals</b>
<ul style="list-style-type: none"> <li>• Hadoop overview.</li> <li>• History.</li> <li>• Hadoop Structure.</li> <li>• HDFS, Hadoop Distributed File System.</li> <li>• Programming Model MapReduce</li> </ul>	<ul style="list-style-type: none"> <li>• Understand and explain the Hadoop suite [Familiarity]</li> <li>• Implement solutions using the MapReduce programming model. [Usage]</li> <li>• Understand how data is saved in the HDFS. [Familiarity]</li> </ul>
<b>Readings:</b> Hwang, Dongarra, and Fox (2011), Buyya, Vecchiola, and Selvi (2013)	

<b>UNIT 3: Procesamiento de Grafos en larga escala (10)</b>	
<b>Competences: a,b,i</b>	
<b>Content</b>	<b>Generales Goals</b>
<ul style="list-style-type: none"> <li>• Pregel: A System for Large-scale Graph Processing.</li> <li>• Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud.</li> <li>• Apache Giraph is an iterative graph processing system built for high scalability.</li> </ul>	<ul style="list-style-type: none"> <li>• Understand and explain the architecture of the Pregel project. [Familiarity]</li> <li>• Understand the GraphLab project architecture. [Familiarity]</li> <li>• Understand the architecture of the Giraph project. [Familiarity]</li> <li>• Implement solutions using Pregel, GraphLab or Giraph. [Usage]</li> </ul>
<b>Readings:</b> Low et al. (2012), Malewicz et al. (2010), Baluja et al. (2008)	

8. Methodology
<p>El profesor del curso presentará clases teóricas de los temas señalados en el programa propiciando la intervención de los alumnos.</p> <p>El profesor del curso presentará demostraciones para fundamentar clases teóricas.</p> <p>El profesor y los alumnos realizarán prácticas</p> <p>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.</p>

## 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.