Peruvian Computing Society (SPC)
School of Computer Science
Sillabus 2021-I

1. COURSE

CS370. Big Data (Mandatory)

2. GENERAL INFORMATION

2.1 Credits	:	3
2.2 Theory Hours	:	1 (Weekly)
2.3 Practice Hours	:	2 (Weekly)
2.4 Duration of the period	:	16 weeks
2.5 Type of course	:	Mandatory
2.6 Modality	:	Face to face
2.7 Prerrequisites	:	• CS272. Data Management II. (5^{th} Sem) • CS3P1. Parallel and Distributed Computing . (8^{th} Sem)

3. PROFESSORS

Meetings after coordination with the professor

4. INTRODUCTION TO THE COURSE

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.

5. 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. COMPETENCES

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

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

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**)

l) Develop principles research in the area of computing with levels of international competitiveness. (**Usage**)

7. SPECIFIC COMPETENCES

a5) Apply efficient techniques to solve computer problems in parallel and distributed environments.

a48) Apply data visualization and/or computer vision and/or GPU programming and/or augmented reality and/or virtual reality to solve problems in our environment.

b4) Identify and efficiently apply various algorithmic strategies and data structures for the solution of a problem given certain space and time constraints.

b5) Identify and efficiently apply diverse algorithmic strategies and data structures for the solution of a problem in parallel and distributed environments.

b6) Implement distributed solutions using MapReduce.

b7) Implement distributed solutions using NoSql databases.

b8) Apply machine learning techniques to large data sets.

b10) Implement distributed solutions using network databases.

i3) Properly use the query optimization, performance, indexing and table fragmentation modules for distributed DBs using an open source database engine such as PostgreSQL, Cassandra or MongoDB

j2) Apply graph and tree theory for optimization and problem solving

l2) Solve problems in our environment based on new proposals for solutions based on computer graphics.

8. TOPICS

Unit 1: Introducción a Big Data (15)	
Competences Expected: a,b,i	
Topics	**Learning Outcomes**
• 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 : [Cou+11]	

Unit 2: Hadoop (15)	
Competences Expected: a,b,i	
Topics	**Learning Outcomes**
• 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 : [HDF11], [BVS13]	

Unit 3: Procesamiento de Grafos en larga escala (10)	
Competences Expected: a,b,i	
Topics	**Learning Outcomes**
• 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+12], [Mal+10], [Bal+08]	

9. WORKPLAN

9.1 Methodology

Individual and team participation is encouraged to present their ideas, motivating them with additional points in the different stages of the course evaluation.

9.2 Theory Sessions

The theory sessions are held in master classes with activities including active learning and roleplay to allow students to internalize the concepts.

9.3 Practical Sessions

The practical sessions are held in class where a series of exercises and/or practical concepts are developed through problem solving, problem solving, specific exercises and/or in application contexts.

10. EVALUATION SYSTEM

********* EVALUATION MISSING ********

11. BASIC BIBLIOGRAPHY

[Bal+08] Shumeet Baluja et al. "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. Beijing, China: ACM, 2008, pp. 895–904. ISBN: 978-1-60558-085-2. DOI: 10.1145/1367497.1367618. URL: http://doi.acm.org/10.1145/1367497.1367618.

[BVS13] Rajkumar Buyya, Christian Vecchiola, and S. Thamarai Selvi. *Mastering Cloud Computing: Foundations and Applications Programming*. 1st. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2013. ISBN: 9780124095397, 9780124114548.

[Cou+11] George Coulouris et al. *Distributed Systems: Concepts and Design*. 5th. USA: Addison-Wesley Publishing Company, 2011. ISBN: 0132143011, 9780132143011.

[HDF11] Kai Hwang, Jack Dongarra, and Geoffrey C. Fox. *Distributed and Cloud Computing: From Parallel Processing to the Internet of Things*. 1st. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN: 0123858801, 9780123858801.

[Low+12] Yucheng Low et al. "Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud". In: *Proc. VLDB Endow.* 5.8 (Apr. 2012), pp. 716–727. ISSN: 2150-8097. DOI: 10.14778/2212351.2212354. URL: http://dx.doi.org/10.14778/2212351.2212354.

[Mal+10] Grzegorz Malewicz et al. "Pregel: A System for Large-scale Graph Processing". In: *ACM SIGMOD Record*. SIGMOD '10 (2010), pp. 135–146. DOI: 10.1145/1807167.1807184. URL: http://doi.acm.org/10.1145/1807167.1807184.