Universidad Nacional Mayor de San Marcos
School of Computer Science
Syllabus of Course
Academic Period 2018-II

1. **Code and Name:** CS3700. Big Data (Mandatory)
2. **Credits:** 3
3. **Hours of theory and Lab:** 1 HT; 4 HL; (15 weeks)
4. **Professor(s)**

   Meetings after coordination with the professor
5. **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`.

6. **Information about the course**

   (a) **Brief description about 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.

   (b) **Prerrequisites:**

   - CS2702. Databases II. (5^{th} Sem)
   - CS3P01. Parallel and Distributed Computing . (7^{th} Sem)

   (c) **Type of Course:** Mandatory

   (d) **Modality:** Face to face

7. **Specific goals of the Course**

   - 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

8. Contribution to 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**)

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**)

9. Competences (IEEE)

C2. Ability to have a critical and creative perspective in identifying and solving problems using computational thinking. ⇒ **Outcome a,b**

C16. Ability to identify advanced computing topics and understanding the frontiers of the discipline.⇒ **Outcome i**

CS2. Identify and analyze criteria and specifications appropriate to specific problems, and plan strategies for their solution.⇒ **Outcome i,b**

CS3. Analyze the extent to which a computer-based system meets the criteria defined for its current use and future development.⇒ **Outcome j**

CS6. Evaluate systems in terms of general quality attributes and possible tradeoffs presented within the given problem.⇒ **Outcome j**

C2. Ability to have a critical and creative perspective in identifying and solving problems using computational thinking. ⇒ **Outcome a,b**

C16. Ability to identify advanced computing topics and understanding the frontiers of the discipline.⇒ **Outcome i**

CS2. Identify and analyze criteria and specifications appropriate to specific problems, and plan strategies for their solution.⇒ **Outcome i,b**

CS3. Analyze the extent to which a computer-based system meets the criteria defined for its current use and future development.⇒ **Outcome j**

CS6. Evaluate systems in terms of general quality attributes and possible tradeoffs presented within the given problem.⇒ **Outcome j**

10. List of topics

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

11. Methodology and Evaluation

Methodology:

Theory Sessions:

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

Lab Sessions:

In order to verify their competences, several activities including active learning and roleplay will be developed during lab sessions.

Oral Presentations:

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

Reading:

Throughout the course different readings are provided, which are evaluated. The average of the notes in the readings is considered as the mark of a qualified practice. The use of the UTEC Online virtual campus allows each student to access the course information, and interact outside the classroom with the teacher and with the other students.

Evaluation System:

12. Content

Unit 1: Introducción a Big Data (15)	
Competences Expected: C2, C4	
Learning Outcomes	**Topics**
• 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]	• Overview on Cloud Computing • Distributed File System Overview • Overview of the MapReduce programming model
Readings : [Cou+11]	

Unit 2: Hadoop (15)	
Competences Expected: C2, C4	
Learning Outcomes	**Topics**
• Understand and explain the Hadoop suite [Familiarity] • Implement solutions using the MapReduce programming model. [Usage] • Understand how data is saved in the HDFS. [Familiarity]	• Hadoop overview. • History. • Hadoop Structure. • HDFS, Hadoop Distributed File System. • Programming Model MapReduce
Readings : [HDF11], [BVS13]	

Unit 3: Procesamiento de Grafos en larga escala (10)	
Competences Expected: C16	
Learning Outcomes	**Topics**
• 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]	• 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.
Readings : [Low+12], [Mal+10], [Bal+08]	