COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI ICADE CIHS

COURSE SYLLABUS 2020-2021

GENERAL INFORMATION

Course information	
Name	Big Data
Code	DTC-MIC-522
Degree	Máster Universitario en Ingeniería Industrial + Máster en Industria Conectada [2nd year]
Semester	2nd (Spring)
ECTS credits	3.0
Type	Compulsory
Department	Telematics and Computer Science
Coordinator	Carlos Miguel Vallez Fernández

Instructor	
Name	Carlos Miguel Vallez Fernández
Department	Telematics and Computing
Office	
e-mail	cmvallez@icai.comillas.edu
Phone	
Office hours	Arrange an appointment through email.
Lab Instructor	
Name	Emilio Martín Gallardo
Department	Telematics and Computing
e-mail	emgallardo@icai.comillas.edu
Office hours	Arrange an appointment through email.

DETAILED INFORMATION

Contextualization of the course

Contribution to the professional profile of the degree

Big data is a new technology that plays a leading role in all processes where there is a large volume of data or where artificial intelligence or machine learning algorithms are required. It is allowing to highly increase efficiency and effectiveness and enabling new business models that were previously impossible or unimaginable. In particular, industry is an area where big data is a key technology, both because artificial intelligence and machine learning techniques result in great process improvements, and because the huge volume of data generated by the Internet of Things (IoT) is otherwise challenging to process and analyze in order to support decision-making.

During the course, students will learn the most relevant aspects of big data technology, go through real cases from leading industrial actors, and carry out some lab practices based on these cases. By the end of the course, students should have enough knowledge of big data technology to understand its potential and have developed an informed criterion to determine when and how to use it in a professional context.

Prerequisites

Students willing to take this course should be familiar with basic probability and statistics, machine learning, and undergraduate-level programming. Previous experience with Python is also desired although not strictly required.

CONTENTS

Contents
Theory
Unit 1. Introduction to Big Data
1.1 What is Big Data and what is not? Origin, motivation and history 1.2 Data value 1.3 Data sources and volumes. Structured and unstructured information 1.4 Life cycle of a Big Data project. Professional profiles: Skills and responsibilities 1.5 The four V's
Unit 2. Distributed systems
2.1 Introduction 2.2 Problems. Fault tolerance. Balancing. Availability. Redundancy 2.3 Linux based operating systems 2.4 Virtualization: Introduction, platforms and virtual machines vs containers
Unit 3. Hadoop ecosystem
3.1 Introduction and components 3.2 Hardware and software architecture 3.3 Administration and monitoring of a Big Data cluster
Unit 4. Massive storage
4.1 HDFS file system 4.2 Roles and services 4.3 HUE
Unit 5. Introduction to massive processing
5.1 YARN 5.2 MapReduce 5.3 Spark
Unit 6. Cloud and Big Data
6.1 On-premise vs Cloud infrastructure 6.2 "As a service" concept 6.3 Cloud platforms and providers
Unit 7. Visualization
7.1 Introduction. Visual encodings, general concepts and history 7.2 Overview of commercial tools 7.3 Use case analyses and best practices
Unit 8. Big Data technology impact in business
8.1 Big Data as an exponential and disruptive technology in business 8.2 Big Data technology use cases

Laboratory
Lab 1. Datacenter tier classification (individual)
Students will become familiar with the datacenter tier classification according to their characteristics.
Lab 2. Server consolidation (group)
Students will have to decide how to distribute several physical servers, which are going to be virtualized, among available target servers.
Lab 3. Virtual machines and containers (group)
Students will learn to run virtual machines and Docker containers.
Lab 4. Linux/Unix commands (individual)
Students will show that they have acquired enough knowledge to work with command line and the terminal of Unix/Linux based computers.
Lab 5. MapReduce (individual)
Students will find a real-life use case that can be addressed with a MapReduce framework.
Lab 6. MapReduce and Hadoop (group)
Students will execute some Hadoop commands and will apply a MapReduce program to real files.
Lab 7. Cloud services (group)
Students will create a comparison table of the most common commercial cloud services with their respective characteristics. The goal is to produce a cheat sheet that allows them to make the decision of which commercial cloud service contract in a real scenario.
Lab 8. Spark (group)
Using Spark code, students will have to answer some questions about a given dataset.

Competences and learning outcomes

Competences[1]

General competences

CG1.	Have acquired advanced knowledge and demonstrated, in a research and technological or highly specialized context, a detailed and well-founded understanding of the theoretical and practical aspects, as well as of the work methodology in one or more fields of study. *Haber adquirido conocimientos avanzados y demostrado, en un contexto de investigación científica y tecnológica o altamente especializado, una comprensión detallada y fundamentada de los aspectos teóricos y prácticos y de la metodología de trabajo en uno o más campos de estudio.*
CG2.	Know how to apply and integrate their knowledge, understanding, scientific rationale, and problem-solving skills to new and imprecisely defined environments, including highly specialized multidisciplinary research and professional contexts. *Saber aplicar e integrar sus conocimientos, la comprensión de estos, su fundamentación científica y sus capacidades de resolución de problemas en entornos nuevos y definidos de forma imprecisa, incluyendo contextos de carácter multidisciplinar tanto investigadores como profesionales altamente especializados.*
CG3.	Know how to evaluate and select the appropriate scientific theory and the precise methodology of their fields of study in order to formulate judgements based on incomplete or limited information, including, when necessary and pertinent, a discussion on the social or ethical responsibility linked to the solution proposed in each case. *Saber evaluar y seleccionar la teoría científica adecuada y la metodología precisa de sus campos de estudio para formular juicios a partir de información incompleta o limitada incluyendo, cuando sea preciso y pertinente, una reflexión sobre la responsabilidad social o ética ligada a la solución que se proponga en cada caso.*
CG4.	Be able to predict and control the evolution of complex situations through the development of new and innovative work methodologies adapted to the scientific/research, technological or specific professional field, in general multidisciplinary, in which they develop their activity. *Ser capaces de predecir y controlar la evolución de situaciones complejas mediante el desarrollo de nuevas e innovadoras metodologías de trabajo adaptadas al ámbito científico/investigador, tecnológico o profesional concreto, en general multidisciplinar, en el que se desarrolle su actividad.*
CG5.	Be able to transmit in a clear and unambiguous manner, to specialist and non-specialist audiences, results from scientific and technological research or state-of-the-art innovation, as well as the most relevant foundations that support them. *Saber transmitir de un modo claro y sin ambigüedades, a un público especializado o no, resultados procedentes de la investigación científica y tecnológica o del ámbito de la innovación más avanzada, así como los fundamentos más relevantes sobre los que se sustentan.*
CG6.	Have developed sufficient autonomy to participate in research projects and scientific or technological collaborations within their thematic area, in interdisciplinary contexts and, where appropriate, with a high knowledge transfer component. *Haber desarrollado la autonomía suficiente para participar en proyectos de investigación y colaboraciones científicas o tecnológicas dentro de su ámbito temático, en contextos interdisciplinares y, en su caso, con una alta componente de transferencia del conocimiento.*
CG7.	Being able to take responsibility for their own professional development and their specialization in one or more fields of study. *Ser capaces de asumir la responsabilidad de su propio desarrollo profesional y de su especialización en uno o más campos de estudio.*

[1] Competences in English are a free translation of the official Spanish version.

Specific competences	
CE5.	Know the techniques used to extract information from large datasets, as well as the different platforms, tools, and languages that make it possible. *Conocer las técnicas para extraer información de grandes conjuntos de datos, así como las diferentes plataformas, herramientas y lenguajes que lo hacen posible.*

Learning outcomes	
By the end of the course students should:	
RA1.	Be able to communicate with big data specialists using common language.
RA2.	List the characteristics and advantages of big data systems, the need that caused its appearance and compare them with alternative solutions.
RA3.	Understand and propose general and industrial applications of big data while being aware of the business impact.
RA4.	Know the state-of-the-art of Big Data, including hands-on experience with several commercial solutions.
RA5.	Be capable of addressing simple analytical projects.
RA6.	Be able to develop analytical algorithms.
RA7.	Properly assess the impact of big data on people, businesses, and society, including ethical and moral aspects.
RA8.	Be able to create small environments to execute proof of concepts based on cloud services, virtual machines of containers.

TEACHING METHODOLOGY

General methodological aspects

To ensure useful and practical learning, theoretical classes will be combined with master classes that reflect the reality of the market. Real case studies with also be studied from business and technical perspectives, some of which will be used in practical sessions.

In-class activities	Competences
▪ **Lectures:** The lecturer will introduce the fundamental concepts of each unit, along with some practical recommendations, and will go through worked examples to support the explanation. Active participation will be encouraged by raising open questions to foster discussion and by proposing quizzes and short application exercises to be solved in class.	CG1, CG3, CG4, CG7, CE5
▪ **Lab sessions:** Under the instructor's supervision, students will apply the concepts learned in the lectures to real cases, in order to face and solve implementation problems that typically arise.	CG1, CG2, CG3, CG5, CG6, CG7, CE5
▪ **Tutoring** for groups or individual students will be organized upon request.	–
Out-of-class activities	**Competences**
▪ Personal study of the course material and resolution of the proposed exercises.	CG1, CG3, CG4, CG7, CE5
▪ Practical session preparation to make the most of in-class time.	CG1
▪ Practical results analysis and report writing.	CG2, CG5, CE5

ASSESSMENT AND GRADING CRITERIA

Assessment activities	Grading criteria	Weight
Quizzes	▪ Understanding of the theoretical concepts. ▪ Class participation	10%
Final exam	▪ Understanding of the theoretical concepts. ▪ Application of these concepts to problem-solving. ▪ Critical analysis of numerical exercises' results.	50%
Individual lab assignments	▪ Application of theoretical concepts to real problem-solving. ▪ Ability to understand results in real environment. ▪ Written communication skills.	25%
Group lab assignments	▪ Application of theoretical concepts to real problem-solving. ▪ Ability to understand results in real environment. ▪ Written communication skills.	15%

GRADING AND COURSE RULES

Grading

Regular assessment

- **Theory** will account for 60%, of which:
  - Quizzes: 10%
  - Final exam: 50%
- **Lab** assignments will account for the remaining 40%, of which:
  - Individual assignments: 25%
  - Group assignments: 15%

In order to pass the course, the weighted average mark must be greater or equal to 5 out of 10 points, the mark of the final exam must be greater or equal to 4 out of 10 points, and the laboratory mark must be at least 5 out of 10 points. Otherwise, the final grade will be the lower of the three marks.

Retake

Lab marks will be preserved as long as they result in a passing grade. Otherwise, a comprehensive lab assignment will have to be developed and handed in. In addition, all students will take a final exam. The resulting grade will be computed as follows:

- **Theory** will account for 60%, of which:
  - Quizzes: 10%
  - Final exam: 50%
- **Lab** will account for the remaining 40%
  - If the student passed the lab during regular assessment
    - Lab assignments: 40%
  - Otherwise
    - Comprehensive lab assignment: 40%

As in the regular assessment period, in order to pass the course, the weighted average mark must be greater or equal to 5 out of 10 points, the mark of the final exam must be greater or equal to 4 out of 10 points, and the mark of the laboratory must be at least 5 out of 10 points. Otherwise, the final grade will be the lower of the three marks.

Course rules
▪ Class attendance is mandatory according to Article 93 of the General Regulations (Reglamento General) of Comillas Pontifical University and Article 6 of the Academic Rules (Normas Académicas) of the ICAI School of Engineering. Not complying with this requirement may have the following consequences: - Students who fail to attend more than 15% of the lectures may be denied the right to take the final exam during the regular assessment period. - Regarding practice, absence to more than 15% of the sessions can result in losing the right to take the final exam of the regular assessment period and the retake. Missed sessions must be made up for credit. ▪ Students who commit an irregularity in any graded activity will receive a mark of zero in the activity and disciplinary procedure will follow (cf. Article 168 of the General Regulations (Reglamento General) of Comillas Pontifical University).

WORK PLAN AND SCHEDULE[2]

In and out-of-class activities	Date/Periodicity	Deadline
Final exam	After the lecture period	–
Lab sessions	From week 3	–
Review and self-study of the concepts covered in the lectures	After each lesson	–
Lab preparation	Before every lab session	–
Lab report writing	–	One week after the end of each session

STUDENT WORK-TIME SUMMARY		
IN-CLASS HOURS		
Lectures	Lab sessions	
21	9	
OUT-OF-CLASS HOURS		
Self-study	Lab assignments and report writing	Video watching
32	24.5	3.5
	ECTS credits:	3 (90 hours)

[2] A detailed work plan of the subject can be found in the course summary sheet (see last page). Nevertheless, this schedule is tentative and may vary to accommodate the rhythm of the class.

BIBLIOGRAPHY

Basic bibliography

- Slides prepared by the lecturer (available in Moodlerooms)
- P. C. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, and G. Lapis, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, 1st Ed., McGraw-Hill, 2012. ISBN-13: 978-0-071-79053-6
- T. White, *Hadoop: The Definitive Guide. Storage and Analysis at Internet Scale*, 4th Ed., O'Reilly Media, 2015. ISBN-13: 978-1-491-90163-2

Complementary bibliography

- L. Joyanes, *Big Data: Análisis de grandes volúmenes de datos en organizaciones* (in Spanish), 1st Ed., Alfaomega, 2013. ISBN-13: 978-8-426-72081-8
- B. Chambers, and M. Zaharia, *Spark: The Definitive Guide. Big Data Processing Made Simple,* 1st Ed., O'Reilly UK, 2018. ISBN-13: 978-1-491-91221-8
- A. Holmes, *Hadoop in Practice*, 2nd Ed., Manning Publications, 2014. ISBN-13: 978-1-617-29222-4
- W. Shotts, *The Linux Command Line: A Complete Introduction*, 5th Ed., No Starch Press, 2019. ISBN-13: 978-1-593-27952-3. [Online]. Available: https://www.linuxcommand.org/tlcl.php
- H. Karau, A. Konwinski, P. Wendell, and M. Zaharia, *Learning Spark: Lightning-Fast Big Data Analysis*, 1st Ed., O'Reilly Media, 2015. ISBN-13: 978-1-449-35862-4
- M. van Steen and A. S. Tanenbaum, *Distributed Systems*, 3rd Ed., CreateSpace Independent Publishing Platform, 2017. ISBN-13: 978-1-543-05738-6. [Online]. Available: https://www.distributed-systems.net/index.php/books/ds3/
- N. Iliinsky and J. Steele, *Designing Data Visualizations*, 1st Ed., O'Reilly Media, 2011. ISBN-13: 978-1-449-31228-2
- Nathan Yau, *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*, 1st Ed., John Wiley & Sons, 2011. ISBN-13: 978-0-470-94488-2
- M. Lutz, *Python Pocket Reference: Python in Your Pocket*, 5th Ed., O'Reilly Media, 2014. ISBN-13: 978-1-449-35701-6
- M. Lutz, *Learning Python: Powerful Object-Oriented Programming*, 5th Ed., O'Reilly Media, 2013. ISBN-13: 978-1-449-35573-9

	In-class activities				Out-of-class activities			
Week	**Time [h]**	**Lecture**	**Laboratory**	**Assessment**	**Time [h]**	**Self-study**	**Lab assignments and report writing**	**Other activities**
1	**2**	Course overview (0.5h) The need for Big Data (1.5h)			**2**	Review and self-study (2h)		
	2	Introduction, history and characteristics of Big Data (1.8h)		Quiz (0.2 h)	**3.5**	Review and self-study (2h)		Film watching (1.5h)
2	**2**	Distributed systems (2h)			**4.5**	Review and self-study (2h)	Datacenter tier classification (2.5h)	
	2	Fault tolerance (2h)			**2**	Review and self-study (2h)		
3	**2**	Virtualization and consolidation (2h)			**4.5**	Review and self-study (2h)	Server consolidation (2.5h)	
	2		Virtual machines and containers (2h)		**6**		Lab preparation (1h) Report writing (3h)	Film watching / tutorials (2h)
4	**2**	Linux/Unix commands (1.8h)		Quiz (0.2 h)	**2**	Review and self-study (2h)		
	2		Linux/Unix commands (2h)		**4**		Lab preparation (1h) Report writing (3h)	
5	**2**	Big data roles and use cases (1h) MapReduce and Hadoop environments (1h)			**4.5**	Review and self-study (2h)	MapReduce (2.5h)	
	2	Hadoop HDFS/YARN (1,8h)		Quiz (0.2 h)	**2**	Review and self-study (2h)		
6	**2**	Cloud and Big Data (1h)	Cloud services (1h)		**3.5**	Review and self-study (2h)	Report writing (1h)	
	2		MapReduce and Hadoop (2h)		**4**		Lab preparation (1h) Report writing (3h)	
7	**2**	Spark (2h)			**2**	Review and self-study (2h)		
	2		Spark (2h)		**4**		Lab preparation (1h) Report writing (3h)	
8	**2**	Visualization (2h)			**2**	Review and self-study (2h)		
				Final exam[3]	**10**	Final exam preparation (10h)		

[3] The final exam will be held on the second week of March.