Syllabus
2024 - 2025

FICHA TÉCNICA DE LA ASIGNATURA

Datos de la asignatura	
Subject name	Big Data Processing Technologies
Subject code	DTC-MBD-515
Mainprogram	N/A
Involved programs	Máster en Big Data. Tec. y Analítica Avanzada/Master in Big Data Technologies and Advanced Analytics [First year]
Level	Master
Quarter	Semestral
Credits	7,5 ECTS
Type	Obligatoria
Department	Department of Telematics and Computer Sciencies

Datos del profesorado	
Teacher	
Name	Patricia Alcalde Sanz
Department	Department of Telematics and Computer Sciencies
EMail	palcalde@icai.comillas.edu
Teacher	
Name	Jorge Ayuso Rejas
Department	Department of Telematics and Computer Sciencies
EMail	jayuso@icai.comillas.edu

DATOS ESPECÍFICOS DE LA ASIGNATURA

Contextualización de la asignatura

Aportación al perfil profesional de la titulación

The purpose of the course is to give an overview of the ingestion and processing tools of the big data environment, especially focused on Spark and Hive.

By the end of the course, students will:

- Be able to choose which is the most appropriate tool to extract data from different sources and take it to a Hadoop cluster.
- Have experience with some processing tools and languages (python, sql, etc).
- Have deep knowledge of spark with python and how to optimize jobs.

Prerrequisitos

Students willing to take this course should be familiar with any programming language, preferably python or SQL and with Linux

commands and utilities.

Competencias - Objetivos

Competencias

Competences

CP1 Integrate architectures, artificial intelligence techniques, advanced data and visualisation analytics and legal compliance to deliver the optimal overall solution.

CP2 Apply and integrate programmatic flows of massive data.

CP4 Implement data processing techniques and use the most common tools appropriate to the conditions and requirements of specific cases.

CP7 Apply advanced knowledge in Big Data and data analytics to develop innovative solutions in projects and research, providing and evaluating optimal solutions for large-scale data processing and analysis.

Skills

HA1 Communicate orally and in writing with technical rigour, clarity of exposition and argumentative coherence to all types of technical and non-technical interlocutors.

HA2 Working in multidisciplinary and/or international teams and organising and leading group dynamics appropriately.

HA3 Developing the interpersonal skills required in today's professional environments (empathy, tolerance, respect, ability to bring together opposing interests).

HA4 Manage, organise and plan work and time appropriately, meeting objectives and quality standards.

HA5 Maintaining continuous training and learning and adapting to technological and scientific changes.

Resultados de Aprendizaje

Knowledge or contents

CO1 Understand the fundamentals of data analytics and its application in different areas of artificial intelligence, highlighting the integration in complex and multidisciplinary solutions for the advanced analysis of massive data attending to the diversity of specific problems in each area.

CO2 Understand the most common and appropriate data processing techniques, architectures and tools for specific case conditions and requirements.

BLOQUES TEMÁTICOS Y CONTENIDOS

Contenidos – Bloques Temáticos

Contents

Theory

Unit 1. Introduction to software development

1. Intellij
2. Git
3. Introduction to SQL

Unit 2. Hadoop Ecosystem

1. Hadoop Ecosystem
2. HDFS and Hadoop client
3. Hive

Unit 3. Processing tools and ETLs

1. Kafka
2. Search Engines
3. NiFi

Unit 4. Introduction to Data scientist

1. Python first steps
2. Scientific Python

Unit 5. Apache Spark for Data scientist

1. Spark DataFrame
2. Spark ML (Machine Learning)
3. Spark packages

Laboratory

All sessions will have a hands-on approach. In the development of the course will be proposed to students practices that will be (of the final grade.

METODOLOGÍA DOCENTE

Aspectos metodológicos generales de la asignatura

EVALUACIÓN Y CRITERIOS DE CALIFICACIÓN

Assessment activities	Grading criteria	Weigh
Practices	▪ Mean of student's practices (0-10 points), all practices must be passed	60%
Final exam	▪ Understanding of the theoretical concepts. ▪ Application of these concepts to problem-solving. ▪ Critical analysis of numerical exercises' results.	40%

BIBLIOGRAFÍA Y RECURSOS

Bibliografía Básica

- Notes and notebooks prepared by the lecturer (available in Moodle).

- White, T. (2015). Hadoop: The definitive guide 4th edition. " O'Reilly Media, Inc.".
- Shreedharan,Hari (2014). Using Flume " O'Reilly Media, Inc."
- Karau, H., Konwinski, A., Wendell, P., & Zaharia, M. (2015). Learning spark: lightning-fast big data analysis. " O'Reilly Media, Inc.".
- VanderPlas, J. (2016). Python Data Science Handbook.