Syllabus
2024 - 2025

TECHNICAL SHEET OF THE SUBJECT

Data of the subject	
Subject name	Big Data Processing Technologies
Subject code	DTC-IMAT-325
Mainprogram	Bachelor's Degree in Mathematical Engineering and Artificial Intelligence
Involved programs	Grado en Ingeniería Matemática e Inteligencia Artificial [Third year]
Credits	6,0 ECTS
Type	Obligatoria (Grado)
Department	Department of Telematics and Computer Sciencies

Teacher Information	
Teacher	
Name	Antonio Javier Samaniego Jurado
Department	Department of Telematics and Computer Sciencies
EMail	ajsamaniego@icai.comillas.edu
Teacher	
Name	Ignacio Pérez Torres
Department	Department of Telematics and Computer Sciencies
EMail	iptorres@icai.comillas.edu
Teacher	
Name	Guillermo Gallego Reina
Department	Department of Telematics and Computer Sciencies
EMail	ggallego@icai.comillas.edu

SPECIFIC DATA OF THE SUBJECT

Contextualization of the subject

Contribution to the professional profile of the degree

Enhance professional profile by providing key skills to effectively manage large datasets, implement distributed systems, automate processes with intelligent agents, visualize data effectively, and stay updated with the latest technological trends. These competencies are essential for standing out in work environments driven by the Big Data era and rapid technological evolution.

Prerequisites

You must understand the operation of distributed systems and the use of virtualized environments.

You must be familiar with the programming techniques acquired in previous courses.

Competencies - Objectives

Competences

GENERALES	
CG04	Conocimientos básicos sobre el uso y programación de los ordenadores, sistemas operativos, bases de datos y programas informáticos con aplicación en ingeniería.
CG05	Conocimiento de la estructura, organización, funcionamiento e interconexión de los sistemas informáticos, los fundamentos de su programación, y su aplicación para la resolución de problemas propios de la ingeniería
CG07	Capacidad para integrarse en equipos de trabajo y colaborar de forma activa con otras personas, áreas y organizaciones en la consecución de los objetivos ligados a las actividades de extracción de valor de los datos e inteligencia artificial.
CG09	Capacidad para determinar eficazmente los objetivos, prioridades, métodos y controles para desempeñar tareas relacionadas con la planificación de proyectos de explotación de datos e inteligencia artificial, mediante la organización de las actividades con los plazos y los medios disponibles
ESPECÍFICAS	
CE14	Dominio de los conceptos y técnicas más utilizadas de adquisición y transformación de la información localizada en local o en remoto en el ámbito del análisis de datos y la inteligencia artificial
CE15	Capacidad para diseñar y gestionar sistemas de almacenamiento de información estructurado, semi-estructurado y no estructurado para el desarrollo de aplicaciones en el ámbito de la inteligencia artificial
CE20	Conocimiento de la infraestructura Big Data de almacenamiento y procesamiento distribuido para el procesamiento de datos masivos
CE21	Capacidad para diseñar e implementar aplicaciones Big Data, siendo capaz de identificar y desplegar las tecnologías que mejor se adapten a cada caso de uso para el procesamiento masivo de datos.
CE23	Capacidad para desarrollar y utilizar herramientas de visualización de grandes volúmenes de datos para poder comunicar los resultados de los análisis realizados sobre los mismos.

Learning outcomes

RA1	Conocer las etapas de un desarrollo de software Big Data así como los perfiles profesionales que intervienen
RA2	Conocer la base de datos noSQL más extendida en el ecosistema Hadoop
RA3	Desarrollar soluciones de ingesta de datos mediante las herramientas más utilizadas en soluciones Big Data
RA4	Conocer los interfaces de consulta pseudo-SQL para entornos Big Data
RA5	Dominar las tecnologías de mensajería de intercambio de información en streaming en soluciones integrales Big Data
RA6	Conocer, diseñar e implementar aplicaciones utilizando el motor de procesamiento más utilizado para plataformas distribuidas: Spark

RA7	Poseer una visión amplia de las tecnologías existentes (tradicionales e innovadoras) en el ecosistema Big Data

THEMATIC BLOCKS AND CONTENTS

Contents - Thematic Blocks

1. Pseudo-SQL Storage
    - Avro.
    - Parquet.

      Raw.
    - JSON.
    - Real Time Data.
    - Others.
2. Big Data Ingestion
    - Apache Nifi.
    - Apache Sqoop.
3. Distributed Message Systems
    - Apache Kafka
4. Storage Systems
    - Hbase.
    - Hive.
    - ElasticSearch.
5. Data Agents
    - Beats.
    - FileBeat.
    - MetricBeat.
6. Data Visualizer
    - Kibanna
7. Big Data Processing Engines
    - SPARK
    - Architecture.
    - Application Design.
    - Performance.
    - Data Visualization.
    - Streaming: SCALA.
    - MLIB.
    - GRAPHX.
8. Model Productivity
    - Flask RestFul Api
9. Technology Trends
    - Generative AI.

TEACHING METHODOLOGY

General methodological aspects of the subject

In-class Methodology: Activities	

The educational activities will take place during the 4 hours of class per week, which will be distributed as follows:
• Expository and participative lectures:
  ◦ The professor will present the theoretical content.
• Practical exercises and problem-solving:
  ◦ The student can ask questions about the theoretical concepts presented in the lecture and the proposed practices.
• Practical sessions using software:
  ◦ Time will be dedicated to understanding and solving these sessions.
• Continuous assessment activities:
  ◦ Tests will be carried out, complementary practices to the weekly ones will be developed, and gamified challenges will be offered.
• Tutoring for doubt resolution:
  ◦ This will be done implicitly during the rest of the described activities.
	CG04, CG05, CG07, CG09, CE14, CE15, CE20, CE21, CE23

Non-Presential Methodology: Activities	

The training activities will be:
• Practical exercises and problem-solving:
  ◦ The students will be given specific problems aimed at understanding the concepts explained in the theoretical sessions. They are required to work on solving these problems remotely, and their solutions should be uploaded to the platform.
• Practical sessions using software:
  ◦ Once the practice is released, the students will work on it remotely.
• Personal study:
  ◦ The main objective of remote work is to understand and comprehend the theoretical concepts of the subject and be able to apply this knowledge to solve different types of problems.
	CG04, CG05, CG07, CG09, CE14, CE15, CE20, CE21, CE23

SUMMARY STUDENT WORKING HOURS

CLASSROOM HOURS			
Clases magistrales expositivas y participativas	Sesiones prácticas con uso de software	Actividades de evaluación continua del rendimiento	Ejercicios prácticos y resolución de problemas
40.00	15.00	2.00	2.00
NON-PRESENTIAL HOURS			
Ejercicios prácticos y resolución de problemas	Sesiones prácticas con uso de software	Estudio personal	Proyectos
3.00	32.00	30.00	50.00
ECTS CREDITS: 6,0 (174,00 hours)			

EVALUATION AND CRITERIA

Evaluation activities	Evaluation criteria	Weight
Final Exam.	• **Final Exam (50%):** **The knowledge acquired in the subject will be evaluated.**	50
Practical sessions: • Collaborative Challenges. • Non-presential work. • Deliverables.	The attitude, participation, and completion of the deliverables and challenges raised in collaborative and individual sessions.	20
Final project	Final project for the subject that the student will submit at the end of the course.	30

Ratings

The final grade in the regular and extraordinary assessment of the subject will depend on the evaluation of the following activities:

Final Grade = 50% Final Exam + 20% Deliverables + 30% Final Project

To pass the subject, students must obtain at least 5 points out of 10 in the final exam of the subject and in the final practice, both in the regular and extraordinary assessment. Absence from 15% or more of the face-to-face hours of this subject may result in the inability to participate in the regular and extraordinary assessments.

BIBLIOGRAPHY AND RESOURCES