



FICHA TÉCNICA DE LA ASIGNATURA

Datos de la asignatura	
Nombre completo	Arquitectura Big Data
Código	DTC-IMAT-312
Título	Grado en Ingeniería Matemática e Inteligencia Artificial
Impartido en	Grado en Ingeniería Matemática e Inteligencia Artificial [Tercer Curso]
Créditos	3,0 ECTS
Carácter	Obligatoria (Grado)
Departamento / Área	Departamento de Telemática y Computación

Datos del profesorado	
Profesor	
Nombre	Antonio Javier Samaniego Jurado
Correo electrónico	ajsamaniego@icai.comillas.edu
Profesor	
Nombre	Ignacio Pérez Torres
Correo electrónico	iptorres@icai.comillas.edu
Profesor	
Nombre	Guillermo Gallego Reina
Departamento / Área	Departamento de Telemática y Computación
Correo electrónico	ggallego@icai.comillas.edu

DATOS ESPECÍFICOS DE LA ASIGNATURA

Contextualización de la asignatura
Aportación al perfil profesional de la titulación
<p>La arquitectura de big data desempeña un papel crucial en el sector de la inteligencia artificial al proporcionar la infraestructura necesaria para manejar volúmenes masivos de datos generados por aplicaciones de IA. Permite el almacenamiento y procesamiento eficiente de datos complejos, como imágenes y texto, esenciales para el entrenamiento y mejora de algoritmos de IA. Además, facilita la implementación de modelos distribuidos, acelerando el procesamiento y análisis de información. Esta integración posibilita un ciclo de retroalimentación donde los resultados de la IA pueden alimentar la mejora continua de los sistemas, impulsando avances significativos en la comprensión y aplicación de la inteligencia artificial.</p>
Prerrequisitos
Debe conocer las técnicas de programación adquiridas en cursos anteriores.

Competencias - Objetivos



Competencias

GENERALES

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.

ESPECÍFICAS

CE13	Conocimiento del funcionamiento, diseño y parametrización de la ejecución de código distribuido en arquitecturas Big Data con el objetivo de crear soluciones de software eficientes para el procesamiento masivo de datos.
CE18	Conocimiento de tecnologías habilitadoras de la transformación digital para el desarrollo de soluciones innovadoras en las organizaciones.
CE20	Conocimiento de la infraestructura Big Data de almacenamiento y procesamiento distribuido para el procesamiento de datos masivos

Resultados de Aprendizaje

RA1	Conocer la necesidad que motivó la aparición de Big data, características, casos de uso y estado del arte del mismo
RA2	Conocer las diferencias entre el despliegue de aplicaciones en entornos stand-alone frente a entornos distribuidos
RA3	Conocer los casos de uso donde es necesario desplegar una infraestructura Big Data
RA4	Conocer los proyectos más relevantes que conforman el ecosistema Hadoop, con sus características, utilidades, funcionalidad y sus opciones alternativas
RA5	Dominar el funcionamiento del sistema de almacenamiento HDFS
RA6	Conocer la historia, evolución, características de YARN y los motores de procesamiento MapReduce y Spark
RA7	Conocer las diferencias entre el despliegue de soluciones Big Data on-premise y los entornos Cloud según el caso de uso

BLOQUES TEMÁTICOS Y CONTENIDOS

Contenidos – Bloques Temáticos

1. Introducción al Big Data
 - Definición y componentes
 - Valor del dato
 - De un entorno local a una arquitectura Big Data
 - Modelos de despliegue
2. Introducción a los sistemas distribuidos en entornos Big Data
 - Introducción



- Fundamentos de los sistemas distribuidos
- Protocolos de red y socket programming
- 3. Arquitecturas avanzadas
 - Introducción
 - Infraestructuras para el despliegue de soluciones Big Data
 - Tecnologías On-Premise
 - Tecnologías Cloud
 - Servicios Big Data cloud
 - Ejemplos de caso de uso de arquitecturas
- 4. Frameworks de computación distribuida
 - Introducción
 - Framework Dask
 - Framework Ray
- 5. Ecosistema Hadoop
 - Introducción
 - HUE. Cloudera Manager
 - Almacenamiento distribuido: HDFS
 - Procesamiento distribuido: YARN. Introducción a MapReduce y Spark

METODOLOGÍA DOCENTE

Aspectos metodológicos generales de la asignatura

Metodología Presencial: Actividades

Las actividades formativas se desarrollarán durante las 2 horas de clase a la semana que se distribuirán:

- **Clases magistrales expositivas y participativas:**
 - Sesión de 2h de Teoría.
 - El profesor realizará un exposición de los contenidos teóricos.
- **Ejercicios prácticos y resolución de problemas:**
 - El alumno planteará dudas sobre los conceptos teóricos expuestos en la clase magistral, y de las prácticas propuestas.
- **Sesiones prácticas con uso de software:**
 - Se dedicará tiempo al entendimiento y resolución de dichas sesiones.
- **Actividades de evaluación continua del rendimiento:** se realizarán pruebas, desarrollarán prácticas complementarias a las semanales y retos gamificados.
- **Tutoría para resolución de dudas:** esta actividad se realizará de forma implícita durante el resto de actividades descritas.

CG05, CG07, CE13, CE18, CE20

Metodología No presencial: Actividades

Las actividades formativas serán:



<ul style="list-style-type: none"> • Ejercicios prácticos y resolución de problemas: <ul style="list-style-type: none"> ◦ El alumno dispondrá de problemas concretos enfocados a asimilar los conceptos explicados en las sesiones teóricas para desarrollar de forma no presencial. La solución de estos problemas será subida a la plataforma. • Sesiones prácticas con uso de software: <ul style="list-style-type: none"> ◦ Una vez liberada la práctica, el alumno trabajará sobre ella de forma no presencial. • Estudio personal: <ul style="list-style-type: none"> ◦ El objetivo principal del trabajo no presencial es llegar a entender y comprender los conceptos teóricos de la asignatura, así como ser capaz de poner en práctica estos conocimientos para resolver los diferentes tipos de problemas. 	CG05, CG07, CE13, CE18, CE20
---	------------------------------

RESUMEN HORAS DE TRABAJO DEL ALUMNO

HORAS PRESENCIALES			
Clases magistrales expositivas y participativas	Sesiones prácticas con uso de software	Actividades de evaluación continua del rendimiento	Ejercicios prácticos y resolución de problemas
14.00	10.00	2.00	2.00
HORAS NO PRESENCIALES			
Trabajos	Estudio personal	Ejercicios prácticos y resolución de problemas	Sesiones prácticas con uso de software
15.00	15.00	3.00	29.00
CRÉDITOS ECTS: 3,0 (90,00 horas)			

EVALUACIÓN Y CRITERIOS DE CALIFICACIÓN

Actividades de evaluación	Criterios de evaluación	Peso
Exámenes: <ul style="list-style-type: none"> • Prueba Intersemestral. • Examen Final. 	<ul style="list-style-type: none"> • Prueba Intersemestral (20%): Comprensión de los conceptos en la introducción a la asignatura • Examen Final (50%): Se evaluará los conocimientos adquiridos en relación a la totalidad de la asignatura <ul style="list-style-type: none"> ◦ Teoría (20%) ◦ Preguntas sobre la elaboración de las prácticas (30%) 	70 %
Sesiones prácticas: <ul style="list-style-type: none"> • Retos Colaborativos. • Trabajos no presenciales. • Prácticas. 	La actitud, participación y realización de las prácticas y los retos planteados en sesiones colaborativas e individuales.	30 %

Calificaciones



La calificación final en convocatoria ordinaria y extraordinaria de la asignatura dependerá de la evaluación de las siguientes actividades:

Nota Final = 20% Prueba Intersemestral + 50% Examen Final + 30% Prácticas semanales

Para aprobar la asignatura los alumnos tienen que obtener al menos 5 puntos sobre 10 en el examen final de la asignatura, tanto en la convocatoria ordinaria como en la extraordinaria.

La inasistencia al 15% o más de las horas presenciales de esta asignatura puede tener como consecuencia la imposibilidad de presentarse a las convocatorias ordinaria y extraordinaria.

BIBLIOGRAFÍA Y RECURSOS

Bibliografía Básica

1. Tom White. Hadoop: The Definitive Guide, 4th Edition (2015). O'Reilly Media, Inc. Libro para ampliar los conocimientos sobre Hadoop.
2. Bill Chambers, Matei Zaharia (2018) Spark: The Definitive Guide. O'Reilly Media, Inc. Libro recomendable para tener unos conocimientos sólidos de Spark.
3. James Warren, Nathan Marz (2015) Big Data: Principles and best practices of scalable realtime data systems. O'Reilly Media, Inc. Libro para conocer los principios y buenas prácticas en los sistemas en tiempo real. Nathan Marz es el creador de Apache Storm y el concepto de Arquitectura Lambda
4. Martin Kleppmann (2017) Designing Data-Intensive Applications. O'Reilly Media, Inc. Martin Kleppmann trabajó en LinkedIn y nos da las claves de diseño de aplicaciones con uso intensivo de datos.
5. GOESSLING, S.; JACKSON, K.L. (2018) Architecting Cloud Computing Solutions. Birmingham: Packt Publishing. Libro para conocer cómo diseñar e implementar soluciones en entornos Cloud de manera efectiva y eficiente.
6. MAXIM, B.; HEISEL, M.; ALI, N. y cols. (2017) Software Architecture for Big Data and the cloud. Burlington: Morgan Kaufmann. Libro para conocer cómo crear arquitecturas en entornos Cloud que den soporte a nuestras aplicaciones Big Data.
7. The Cloud Data Lake: A Guide to Building Robust Cloud Data Architecture
8. The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science
9. DAMA-DMBOK: Data Management Body of Knowledge: 2nd Edition
10. Ramcharan Kakarla, Sundar Krishnan, Sridhar Alla - Applied Data Science Using PySpark_ Learn the End-to-End Predictive Model-Building Cycle (2021, Apress)

En cumplimiento de la normativa vigente en materia de **protección de datos de carácter personal**, le informamos y recordamos que puede consultar los aspectos relativos a privacidad y protección de datos que ha aceptado en su matrícula entrando en esta web y pulsando "descargar"

<https://servicios.upcomillas.es/sedelectronica/inicio.aspx?csv=02E4557CAA66F4A81663AD10CED66792>