



GUÍA DOCENTE

PROCESAMIENTO DE DATOS

GRADO EN INGENIERÍA DEL SOFTWARE

MODALIDAD: PRESENCIAL

CURSO ACADÉMICO: 2024-2025

Denominación de la asignatura:	Procesamiento de Datos
Titulación:	Ingeniería del Software
Facultad o Centro:	Centro Universitario de Tecnología y Arte Digital
Materia:	Ingeniería de Datos
Curso:	3º
Cuatrimestre:	2
Carácter:	OBM
Créditos ECTS:	6
Modalidad/es de enseñanza:	Presencial
Idioma:	Castellano
Profesor/a - email	Miguel Angel Fernandez Diaz /miguel.fernandez@u-tad.com
Página Web:	http://www.u-tad.com/

DESCRIPCIÓN DE LA ASIGNATURA

Descripción de la materia

Los contenidos de la materia permiten a los alumnos comprender el flujo de búsqueda, ingesta, almacenamiento, procesamiento y análisis de información de datos y aproxima a los alumnos a las técnicas y tecnologías necesarias para la gestión de grandes cant

Descripción de la asignatura

El objetivo de este curso es entrar en el mundo de Big Data en un entorno distribuido y en real time.

Aprender procesamiento de datos distribuido es algo fundamental hoy en día, en este curso utilizaremos Spark. Spark es la tecnología que está revolucionando el mundo de la analítica y el big data. Spark es un motor de procesamiento de datos de código abierto creado en torno a la velocidad, la facilidad de uso y el análisis.

Empezaremos el curso con una Introducción a Scala con el objetivo de utilizar Spark usando SparkShell. Veremos conceptos básicos de Spark como tareas distribuidas, RDD y al arquitectura Máster/Slave y a continuación exploraremos los módulos/extensiones Spark SQL, Spark Streaming y Machine Learning con MLib en Spark.

COMPETENCIAS Y RESULTADOS DE APRENDIZAJE

Competencias (genéricas, específicas y transversales)

COMPETENCIAS BÁSICAS Y GENERALES

CB1: Que los estudiantes hayan demostrado poseer y comprender conocimientos en un área de estudio que parte de la base de la educación secundaria general, y se suele encontrar a un nivel que, si bien se apoya en libros de texto avanzados, incluye también algunos aspectos que implican conocimientos procedentes de la vanguardia de su campo de estudio.

CB2: Que los estudiantes sepan aplicar sus conocimientos a su trabajo o vocación de una forma profesional y posean las competencias que suelen demostrarse por medio de la elaboración y defensa de argumentos y la resolución de problemas dentro de su área de estudio.

CB3: Que los estudiantes tengan la capacidad de reunir e interpretar datos relevantes (normalmente dentro de su área de estudio) para emitir juicios que incluyan una reflexión sobre temas relevantes de índole social, científica o ética.

CB4: Que los estudiantes puedan transmitir información, ideas, problemas y soluciones a un público tanto especializado como no especializado.

CB5: Que los estudiantes hayan desarrollado aquellas habilidades de aprendizaje necesarias para emprender estudios posteriores con un alto grado de autonomía

CG1 - Capacidad para entender, planificar y resolver problemas a través del desarrollo de soluciones informáticas.

CG3 - Conocimiento de los fundamentos científicos aplicables a la resolución de problemas informáticos

CG4 - Capacidad para simplificar y optimizar los sistemas informáticos atendiendo a la comprensión de su complejidad

CG9 - Capacidad para aprender, modificar y producir nuevas tecnologías informáticas

CG10 - Uso de técnicas creativas para la realización de proyectos informáticos

CG11 - Capacidad de buscar, analizar y gestionar la información para poder extraer conocimiento de la misma

COMPETENCIAS ESPECÍFICAS

CE3 - Conocimiento del álgebra relacional y realización de consultas en lenguajes procedurales para el diseño de esquemas de

bases de datos normalizados basados en modelos de entidad-relación

CE10 - Capacidad para manejar un gestor de versiones de código y generar la documentación de una aplicación de forma

automática.

COMPETENCIAS TRANSVERSALES

CT1 - Conocimiento de la definición, el alcance y la puesta en práctica de los fundamentos de las metodologías de gestión de proyectos de desarrollo tecnológico

CT2 - Conocimiento de los principales agentes del sector y del ciclo de vida completo de un proyecto de desarrollo y comercialización de contenidos digitales

CT4 - Capacidad de actualización del conocimiento adquirido en el manejo de herramientas y tecnologías digitales en función del estado actual del sector y de las tecnologías empleadas

CT5 - Desarrollo de las habilidades necesarias para el emprendimiento digital.

Resultados de aprendizaje

Al acabar la titulación, el graduado o graduada será capaz de:

- Comprender e implementar los métodos de almacenamiento y administración eficaz en entornos distribuidos de datos no estructurados.
- Conocer y saber aplicar las distintas técnicas de aprendizaje supervisado, semi-supervisado y no supervisado.
- Entender y aplicar las técnicas de Deep learning
- Ser capaz de recuperar información mediante técnicas de web scraping o APIs normalizadas
- Entender y aplicar las técnicas de análisis del lenguaje natural
- Ser capaz de analizar contenidos de redes sociales
- Entender la naturaleza y representación de las imágenes digitales.
- Conocer las aplicaciones de las redes neuronales al análisis y generación de sonido, imagen estática y video.
- Desarrollar soluciones informáticas aplicadas a la visión por computador.
- Desarrollar un proyecto completo de datos aplicando metodología iterativa, desde el diseño hasta el despliegue.

CONTENIDO

Fragmentación y distribución de datos.

Concurrencia distribuida.

Protocolos de confiabilidad y confirmación.

Administración de datos replicados.

Arquitecturas distribuidas de gestión de datos.

TEMARIO

Tema 1.- Introducción a Spark y Scala

Introducción y instalación de Spark. Introducción a las estructuras de datos en Scala: listas, diccionarios y data frames. Métodos de manipulación de estructuras: lista comprendida, funciones anónimas/lambda y vía map/reduce/filter, apply y fold.

Tema 2.- Spark Básico

Conceptos básicos de Spark. Spark Core, tareas distribuidas, programación y funcionalidades básicas de I/O y RDD (Resilient Distributed Datasets).

Tema 3.- Spark Cluster

Arquitectura Máster/Slave - esclavos/trabajadores en el caso de Spark. El controlador y los ejecutores ejecutan sus procesos individuales y los usuarios pueden ejecutarlos en el mismo clúster de Spark o en máquinas separadas.

Tema 4.- Spark SQL

Spark SQL es un módulo Spark para el procesamiento de datos estructurados. Proporciona una abstracción de programación llamada DataFrames y también puede actuar como un motor de consultas SQL distribuido.

Tema 5.- Spark Streaming

Spark Streaming es una extensión de la API principal de Spark que permite el procesamiento de flujos escalable, de alto rendimiento y tolerante a fallas de flujos de datos en tiempo real. Spark Streaming proporciona una abstracción de alto nivel llamada flujo discretizado o DStream, que representa un flujo continuo de datos.

Tema 6.- Machine Learning con MLib en Spark

MLlib es la biblioteca de aprendizaje automático (ML) de Spark. Su objetivo es hacer que el aprendizaje automático práctico sea escalable y fácil. A un alto nivel, proporciona herramientas como: Algoritmos ML: algoritmos de aprendizaje comunes como clasificación, regresión, agrupamiento y filtrado colaborativo.

ACTIVIDADES FORMATIVAS Y METODOLOGÍAS DOCENTES

Actividades formativas

Actividad Formativa	Horas totales	Horas presenciales
<i>Clases teóricas / Expositivas</i>	29,38	29,38
<i>Clases Prácticas</i>	23,25	23,25
<i>Tutorías</i>	4,00	2
<i>Estudio independiente y trabajo autónomo del alumno</i>	50,00	0,00
<i>Elaboración de trabajos (en grupo o individuales)</i>	31,88	0,00

<i>Actividades de Evaluación</i>	5,25	5,25
<i>TOTAL</i>	143,76	59,88

Metodologías docentes

Método expositivo o lección magistral

Aprendizaje de casos

Aprendizaje basado en la resolución de problemas

Aprendizaje basado en proyectos

Aprendizaje cooperativo o colaborativo

Aprendizaje por indagación

Metodología Flipped classroom o aula invertida

Gamificación

Just in time Teaching (JITT) o aula a tiempo

Método expositivo o lección magistral

Método del caso

Aprendizaje basado en la resolución de problemas

Aprendizaje basado en proyectos

Aprendizaje cooperativo o colaborativo

Aprendizaje por indagación

Metodología flipped classroom o aula invertida

Gamificación

DESARROLLO TEMPORAL

UNIDADES DIDÁCTICAS / TEMAS

PERÍODO TEMPORAL

Introducción a la asignatura, instalación de máquina virtual Semana 1

Introducción a scala (estructuras y funciones, lambda y map/reduce/filter) Semana 2 a 3

Spark Básico Semanas 4 a 6

Spark Cluster Semanas 7 a 8

Spark SQL Semanas 9 a 10

Parcial Semanas 11

Spark Streaming Semanas 12 a 13

Machine Learning con MLib en Spark Semanas 14 a 15

SISTEMA DE EVALUACIÓN

ACTIVIDAD DE EVALUACIÓN	VALORACIÓN MÍNIMA RESPECTO A LA CALIFICACIÓN FINAL (%)	VALORACIÓN MÁXIMA RESPECTO A LA CALIFICACIÓN FINAL (%)
<i>Evaluación de la participación en clase, en prácticas o en proyectos de la asignatura</i>	10	30
<i>Evaluación de trabajos, proyectos, informes, memorias</i>	40	80
<i>Prueba Objetiva</i>	10	60

CRITERIOS ESPECÍFICOS DE EVALUACIÓN

ACTIVIDAD DE EVALUACIÓN	CONVOCATORIA ORDINARIA	CONVOCATORIA EXTRAORDINARIA
<i>Evaluación de la participación en clase, en prácticas o en proyectos de la asignatura</i>	10	10
<i>Evaluación de trabajos, proyectos, informes, memorias</i>	40	40
<i>Prueba Objetiva</i>	50	50

Consideraciones generales acerca de la evaluación

- Drinks and food are not allowed in the classroom. Drinks of any kind are not allowed on the tables, even in sealed containers.

BIBLIOGRAFÍA / WEBGRAFÍA

Bibliografía Básica:

Holden Karau, , Learning Spark: Lightning-Fast Big Data, 2015

Bibliografía Recomendada:

Libro online de IA y Big Data: <https://iaarbook.github.io/>

MATERIALES, SOFTWARE Y HERRAMIENTAS NECESARIAS

Tipología del aula

Aula teórica

Equipo de proyección y pizarra

Materiales:

Ordenador personal

Software:

Virtual Box

Ubuntu 22.04