FICHA IDENTIFICATIVA

Datos de la Asignatura

Código	36437
Nombre	Datos masivos
Ciclo	Grado
Créditos ECTS	6.0
Curso académico	2023 - 2024

Titulación(es)

Titulación	Centro	Curso	Periodo
1406 - Grado en Ciencia de Datos	Escuela Técnica Superior de Ingeniería	4	Primer cuatrimestre

Materias

Titulación	Materia	Caracter
1406 - Grado en Ciencia de Datos	12 - Computación	Obligatoria

Coordinación

Nombre	Departamento
LIBEROS MASCARELL, ALEJANDRO	242 - Ingeniería Electrónica
RODRIGO BORT, MIGUEL	242 - Ingeniería Electrónica

RESUMEN

La asignatura “Datos masivos” es una asignatura cuatrimestral, consta de 6 créditos ECTS y se imparte durante el 1er cuatrimestre del cuarto curso del Grado en Ciencia de Datos de la Universitat de València.

Esta asignatura pretende consolidar las competencias del alumnado en la adquisición, almacenamiento y tratamiento de datos, y a la vez aportar las herramientas para la escalabilidad de estas técnicas al volumen de datos masivos.

De manera general, al finalizar el curso el alumnado debe ser capaz de:
- Entender el concepto de datos masivos o Big Data, conocer su impacto social y aplicaciones.
- Conocer y saber utilizar los diferentes recursos hardware y software para el almacenamiento de datos.
- Conocer y saber utilizar las herramientas hardware y software propias de los sistemas de ficheros distribuidos.
- Conocer y saber utilizar técnicas básicas de programación paralela para el acceso y tratamiento de sistemas de datos distribuidos.
- Conocer y saber desarrollar diferentes algoritmos avanzados para la explotación de datos masivos haciendo uso de las capacidades de computación distribuida de las arquitecturas introducidas.

La información de contacto con el profesorado responsable está publicada en la web del Departamento de Ingeniería Electrónica (http://www.uv.es/die). El material de la asignatura (apuntes, guiones de prácticas, actividades, etc.) estará al alcance del alumnado a través de Aula Virtual, la plataforma de e-learning de la Universitat de València (http://aulavirtual.uv.es/).

El profesorado hará uso, preferiblemente, del correo electrónico para convocar actos de evaluación, puntualizaciones sobre esta guía docente y otros aspectos relevantes de cara al proceso de enseñanza-aprendizaje. De igual manera, el profesorado podrá introducir tareas de evaluación continua u otros aspectos relevantes en el desarrollo diario al aula, así como en el Aula Virtual de la asignatura.

Las clases de teoría se impartirán en castellano y las clases prácticas y de laboratorio según consta en la ficha de la asignatura disponible en la web del grado.

CONOCIMIENTOS PREVIOS

Relación con otras asignaturas de la misma titulación

No se han especificado restricciones de matrícula con otras asignaturas del plan de estudios.

Otros tipos de requisitos

La asignatura se desarrolla en el último curso del Grado en Ciencia de Datos, y por tanto supone una etapa de consolidación y ampliación de gran parte de los conocimientos trabajados durante el grado.
Sin haber requisitos previos para la matriculación de la asignatura, se recomienda haber cursado previamente las asignaturas de Redes y Seguridad, Programación Paralela e Infraestructura de Almacenamiento de Datos. Se recomienda también haber trabajado previamente con los lenguajes de programación trabajado

COMPETENCIAS

1406 - Grado en Ciencia de Datos

- (CG02) Capacidad de resolver problemas con iniciativa, creatividad, y de comunicar y transmitir conocimientos, habilidades y destrezas, comprendiendo la responsabilidad ética y profesional de la actividad del Científico de Datos.
- (CT03) Habilidad para defender su trabajo con rigor y argumentos, exponiéndolo de forma adecuada y precisa, apoyándose en los medios necesarios.
- (CT05) Capacidad para evaluar las ventajas e inconvenientes de diferentes alternativas metodológicas y/o tecnológicas en distintos ámbitos de aplicación.
- (CE02) Conocer y aplicar de forma metodológica las técnicas de programación y la algoritmia necesarias para el procesado eficiente de información y la resolución informática de problemas que utilizan grandes volúmenes de datos.

- (CE04) Conocer y utilizar los distintos modelos de almacenamiento de datos y los sistemas de gestión de las bases de datos utilizando lenguajes de programación de definición, consulta y manipulación de los mismos.
- (CE08) Capacidad para comprender, seleccionar y utilizar la infraestructura y técnicas adecuadas para el tratamiento de datos masivos, atendiendo a criterios de eficiencia, escalabilidad, seguridad, tolerancia a fallos y adecuación al entorno de producción.
- (CE11) Capacidad para diseñar e implementar la toma de datos, su integración, transformación, selección, comprobación de su calidad y veracidad a partir de distintas fuentes, teniendo en cuenta su carácter, heterogeneidad y variabilidad.
- (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.
- (CB5) Que los estudiantes hayan desarrollado aquellas habilidades de aprendizaje necesarias para emprender estudios posteriores con un alto grado de autonomía.

RESULTADOS DE APRENDIZAJE

Los objetivos generales de la asignatura descritos en el resumen junto con las competencias generales, básicas, transversales y específicas, resultan en los siguientes resultados de aprendizaje (RA).

Identificar y describir los requerimientos de almacenamiento y procesamiento de los sistemas de datos masivos.
RA1- Conocer y saber utilizar los servicios y herramientas ofrecidos por los sistemas operativos para el almacenamiento de datos (CG02, CB5, CE04).
RA2- Conocer y saber utilizar los diferentes niveles de almacenamiento local, desde los dispositivos físicos de almacenamiento hasta los sistemas de ficheros (CG03, CB2, CE08, CE11).
RA4- Conocer la estructura básica de soporte físico usado en "Big Data" (CG03, CB2, CE08, CE11, CT05).
RA9- Identificar y describir los requerimientos de almacenamiento y procesamiento de los sistemas de datos masivos (CG03, CB2, CT05, CE04).

Describir y aplicar los nuevos paradigmas de procesamiento paralelo y distribuido necesarios en los sistemas de datos masivos.
RA5- Conocer la estructura de capas de las redes de computadores, así como los principales protocolos y servicios usados en Internet y en el tratamiento de datos (CG03, CB5, CE04).
RA6- Conocer y saber utilizar los dispositivos físicos y virtuales necesarios para crear y mantener redes de procesadores (CG02, CB5, CT03, CE02).
RA10- Describir y aplicar los nuevos paradigmas de procesamiento paralelo y distribuido necesarios en los sistemas de datos masivos (CG03, CB2, CT03, CE02).

Conocer la arquitectura y gestionar los recursos de sistemas de ficheros distribuidos.
RA3- Conocer las técnicas habituales de virtualización de recursos y saber seleccionarlas y utilizarlas (CG03, CB2, CT05, CE11).
RA7- Conocer los riesgos derivados de la obtención, procesamiento, almacenamiento e intercambio de datos (CG02, CB2, CT03, CE04).
RA11- Conocer la arquitectura y gestionar los recursos de sistemas de ficheros distribuidos (CG02, CB2,

CT03, CE08).

Conocer y aplicar diferentes paradigmas sobre arquitectura, flujo de datos y el modelo de programación de datos masivos.
RA8- Seleccionar y aplicar las medidas técnicas que permitan mantener la seguridad de los sistemas de obtención, procesamiento, almacenamiento e intercambio de datos (CG02, CB2, CT03, CE08).
RA12- Conocer y aplicar diferentes paradigmas sobre arquitectura, flujo de datos y el modelo de programación de datos masivos (CE02, CT03,CB5, CG02).
RA16- Evaluar las prestaciones y escalabilidad de un sistema de procesamiento paralelo, estableciendo y aplicando las métricas para su comparación (CG03, CB2, CT03, CE02).

Conocer y utilizar los servicios ofrecidos por los sistemas de computación en la nube para el procesamiento de datos masivos.
RA13- Conocer y utilizar los servicios ofrecidos por los sistemas de computación en la nube para el procesamiento de datos masivos (CG02, CB2, CT05, CE02).
RA14- Identificar y describir las arquitecturas de los computadores paralelos y distribuidos (CE04, CT03, CB2, CG03).
RA15- Conocer y aplicar los paradigmas de programación paralela y distribuida, los modelos de programación relacionados y los estándares para el desarrollo de sistemas de altas prestaciones (CG03, CB2, CT03, CE02).
RA17- Diseñar y desarrollar algoritmos distribuidos que exploten las capacidades de paralelismo de las infraestructuras de computación paralela y distribuida (CE02, CT03, CB5, CG02).
RA18- Diseñar y desarrollar programas que utilicen con eficiencia los multiprocesadores y las arquitecturas distribuidas para el procesado de datos (CE02, CT03, CB5, CG02).

DESCRIPCIÓN DE CONTENIDOS

1. Introducción a los datos masivos

En esta unidad se definirá el concepto de datos masivos o Big Data, la dependencia de esta definición con los recursos hardware disponibles/necesarios. También se tratará el impacto de esta tecnología en los últimos años. (RA4)

2. Almacenamiento de datos

En esta unidad se profundizará en el almacenamiento de datos desde el punto de vista de los recursos software y hardware (físicos y virtuales) que permiten la definición, consulta, mantenimiento y manipulación de volúmenes de datos masivos. Se introducirá y trabajará el modelo de almacenamiento de datos de HDFS (Hadoop Distributed File System). (RA: 1-4).

3. Datos distribuidos

En esta unidad se trabajará sobre las herramientas hardware y software para la generación, mantenimiento, acceso y tratamiento de bases de datos distribuidas en redes interconectadas. Se prestará especial atención a los protocolos que aseguran la integridad de los datos en los sistemas de ficheros distribuidos y a la virtualización de redes o sistemas de computación en la nube, entre otras. Se trabajará con el protocolo HDFS así como con sistemas de almacenamiento distribuido en plataformas

en la nube como GoogleCloud (RA: 4-9, 11, 13).

4. Programación distribuida

En esta unidad se introducirán las técnicas básicas para el tratamiento y procesado de datos de manera distribuida/paralela en redes de procesadores. Se prestará especial atención al modelo de programación MapReduce y al entorno Apache Spark. (RA: 5-6, 8-15).

5. Aplicaciones y algoritmos de programación distribuida.

En esta unidad se consolidarán las competencias del alumnado en la explotación de los datos masivos mediante el diseño y evaluación de algoritmos de programación distribuida de alto nivel que exploten las capacidades de paralelismo de las infraestructuras de computación paralela. Se trabajará en el aprendizaje de Apache Spark sobre bases de datos distribuidas con HDFS en plataformas en la nube como GoogleCloud (RA: 10-18).

VOLUMEN DE TRABAJO

ACTIVIDAD	Horas	% Presencial
Clases de teoría	34,00	100
Prácticas en laboratorio	20,00	100
Prácticas en aula	6,00	100
Elaboración de trabajos en grupo	20,00	0
Estudio y trabajo autónomo	24,00	0
Lecturas de material complementario	6,00	0
Preparación de actividades de evaluación	24,00	0
Preparación de clases de teoría	6,00	0
Preparación de clases prácticas y de problemas	10,00	0
TOTAL	**150,00**	

METODOLOGÍA DOCENTE

Clases presenciales/síncronas
En las clases se desarrollarán los contenidos de la materia empleando una metodología expositiva. Se instará también la participación del alumnado a través de diferentes medios: se podrán emplear cuestionarios on-line, cuestiones abiertas o cualquier otra herramienta para evaluar el grado de consolidación de la materia.
Esta metodología expositiva/interactiva se alternará con prácticas de aula en las que se resolverán casos prácticos o se estudiarán casos de aplicación entre otros.
Competencias más relevantes: CE02, CE04, CE08, CE11, CT01, CT02, CT03, CB1, CB2, CB3, CB4, CB5, CG01, CG02, CG04, CG05, CG07.

Prácticas de laboratorio
Los contenidos trabajados en clase se consolidarán con diferentes prácticas de laboratorio de asistencia obligatoria donde se propondrán diferentes ejercicios prácticos relacionados con la materia. Los resultados de estas prácticas se entregarán de forma escalonada a lo largo del curso para su evaluación, de acuerdo con las indicaciones del profesorado.
Competencias más relevantes: CE02, CE04, CE08, CE11, CT01, CT02, CT03, CT05, CB2, CB4, CG02, CG03, CG04, CG05, CG07.

Preparación de trabajos prácticos y teóricos
Además de las tareas descritas, el alumnado deberá realizar tareas de manera no presencial asociadas a la preparación de prácticas y la elaboración de memorias e informes en los que se debe prestar especial atención a la descripción con rigor y con capacidad crítica de las diferentes metodologías empleadas y decisiones tomadas (CT03, CT05). Además, será imprescindible un estudio individual de la materia tanto de manera continua como de cara a los diferentes actos de evaluación. Durante el curso se podrá además solicitar la consulta de publicaciones o la asistencia a charlas de interés relacionadas con la materia.
Competencias más relevantes: CE02, CE04, CE08, CE11, CT03, CT05, CB2, CB3, CB4, CG02, CG03, CG04, CG05, CG07.

Tutorías
A lo largo del curso el profesorado atenderá a dudas tanto a través de e-mail, como de tutorías en formato presencial o por teleconferencia. La forma de contacto con el profesorado será el correo electrónico tal y como se indicará en Aula Virtual. Además, la información de contacto con el profesorado responsable está publicada en la web del Departamento de Ingeniería Electrónica (http://www.uv.es/die).

Plataforma e-learning y comunicación
El material de la asignatura (apuntes, guiones de prácticas, actividades, etc.) estará al alcance del alumnado a través de Aula Virtual, la plataforma de e-learning de la Universitat de València (http://aulavirtual.uv.es/).

El profesorado hará uso, preferiblemente, del correo electrónico para convocar actos de evaluación, puntualizaciones sobre esta guía docente y otros aspectos relevantes de cara al proceso de enseñanza-aprendizaje. De igual manera, el profesorado podrá introducir tareas de evaluación continua u otros aspectos relevantes en el desarrollo diario al aula, así como en el Aula Virtual de la asignatura.

EVALUACIÓN

En lo que respecta a la evaluación se tendrán en cuenta diferentes dimensiones del proceso de enseñanza-aprendizaje. En primer lugar, el sistema de evaluación responde a las diferentes competencias, resultados de aprendizaje y contenidos a trabajar durante el curso. En segundo lugar, se balancearán tanto aquellas actividades desarrolladas en grupo como el trabajo individual. Finalmente, la evaluación se propone como formativa, es decir, se facilitarán comentarios que favorezcan la subsanación de aspectos a mejorar detectados durante el curso, ya sea en la interacción diaria entre alumnado y profesorado, a través de comentarios en Aula Virtual o en sesiones de revisión.

Tanto en primera como en segunda convocatoria, la nota final (NF) responde a los diferentes sistemas de evaluación (SE) atendiendo a la siguiente expresión:

$$NF = SE1 \cdot 0,5 + SE2 \cdot 0,3 + SE3 \cdot 0,2$$

SE1: Prueba objetiva individual. (Instrumentos de evaluación: Ex1/Ex2)
SE2: Evaluación de prácticas. (Instrumentos de evaluación: Labs/LabEx)
SE3: Evaluación continua. (Instrumentos de evaluación: EjP/TC/otros)

En cualquier caso: (1) si la nota SE1 tiene un valor menor a 5/10 laNF será igual a SE1; (2) si la nota SE2 tiene un valor menor a 5/10, NF será igual a SE2; (3) NF debe ser superior en primera o segunda convocatoria a 5/10 para superar la asignatura

A continuación, se describen los diferentes instrumentos de evaluación:

Ex1/Ex2: Examen individual (SE1). Podrá contener tanto cuestiones breves, como de desarrollo de cuestiones teórico-prácticas, problemas, casos de estudio etc. Se podrá preguntar sobre cualquier aspecto trabajado durante el curso, también podrán aparecer nuevos problemas relacionados con la materia, al considerarse esta una metodología útil para valorar la consolidación de las competencias y contenidos. Esta prueba se realizará de acuerdo con el calendario de exámenes de la escuela, Ex1 corresponde a la primera convocatoria y Ex2 a la segunda.
La participación en Ex2 será obligatoria siempre que no se supere la asignatura en primera convocatoria, en caso contrario la nota en segunda convocatoria será de No Presentado, cualquier excepción a este respecto deberá ser autorizada por el profesorado.
Competencias más relevantes: CE02, CE04, CE08, CE11, CT03, CT05, CB2, CB3, CB4, CG02, CG07.

Labs: Laboratorios (SE2). Durante el curso, y preferiblemente en equipos de dos personas, se realizarán prácticas de laboratorio. Se evaluará tanto el desarrollo en el aula, como los informes asociados a cada práctica que deberán discutir los procedimientos empleados con rigor y en su caso responder a las preguntas planteadas. Además, se podrán solicitar tareas de preparación de la práctica que podrán ser igualmente evaluables hasta un máximo de un tercio de la nota correspondiente a cada práctica. Para tener una nota asociada a cada práctica será obligatoria la asistencia.
En primera convocatoria, SE2 estará determinada por la evaluación continua de los laboratorios a través de los métodos descritos anteriormente.
La copia en cualquiera de estas actividades será penalizada de manera estricta pudiéndose anular todas las notas de evaluación de laboratorio. La no asistencia de manera no justificada y reiterada a las sesiones de laboratorio supondrá una nota de laboratorio en primera convocatoria de 0.
Competencias más relevantes: CE02, CE04, CE08, CE11, CT01, CT02, CT03, CT05, CB2, CB4, CG02, CG03, CG04, CG05, CG07.

LabEx: Examen de laboratorio (SE2). El alumnado que no supere en primera convocatoria la calificación mínima en SE2 a través del método descrito anteriormente, deberá realizar un examen de laboratorio. En este examen se evaluará su desempeño con las herramientas utilizadas durante el curso y su capacidad para interpretar resultados, entre otras competencias asociadas a los laboratorios. Este examen se realizará en segunda convocatoria y de acuerdo con el calendario oficial.
Competencias más relevantes: CE02, CE04, CE08, CE11, CT03, CT05, CB2, CB4, CG02, CG03, CG07.

EjP: Ejercicio parcial individual (SE3). Durante el curso y en horario de clase, se realizará una prueba para evaluar la consolidación de contenidos y competencias, así como dar la oportunidad al alumnado de enfrentarse a ejercicios similares a los que se podrá encontrar en Ex1/Ex2. Los contenidos aplicables a dicha prueba, así como las siguientes normas a seguir y la fecha se comunicará durante el curso. En ningún caso esta prueba eliminará materia de cara a Ex1/Ex2. El peso de este ejercicio será de un 10% de

la nota final.
Competencias más relevantes: CE04, CE08, CE11, CT03, CT05, CB2, CB3, CB4, CG02, CG07.

TC: Tareas colaborativas (SE3). La evaluación continua se verá completada a través de distintas tareas colaborativas que se podrán plantear tanto para su realización en el aula como de forma no presencial. Estas tareas se realizarán en equipo, se trabajará entre otras la coordinación entre diferentes miembros de un equipo, la discusión para alcanzar soluciones de consenso, el análisis de bibliografía relevante, etc. Se podrán emplear técnicas de evaluación continua y por pares para diferenciar las notas de diferentes miembros de un equipo. El peso de este apartado será de un 10% de la nota final.
No se tendrán en cuenta actividades o memorias entregadas fuera de plazo, ni se podrán recuperar actividades no realizadas.
Competencias más relevantes: CE02, CE04, CE08, CE11, CT01, CT02, CT03, CB1, CB2, CB3, CB4, CB5, CG01, CG02, CG04, CG05, CG07.

Otros (SE3): En cualquier caso, el profesorado se reserva la posibilidad de añadir otros métodos de evaluación continua como: observación diaria, controles de asistencia y participación… que podrán sustituir/complementar al apartado TC.

En segunda convocatoria, y siguiendo las notas mínimas indicadas anteriormente, se podrá calcular la nota final como: NF = SE1·0,6+SE2·0,4

En cualquier caso, el sistema de evaluación se regirá por lo que establece el Reglamento de Evaluación y Calificación de la Universitat de Valencia para Grados y Másters.

REFERENCIAS

Básicas

- Large Scale Machine Learning with Python, B. Sjardin, A. Boscheti, L. Massaron.
- Big Data: Principles and Best Practices. N. Marz, J. Warren
- Spark, the definitive guide, B. Chambers, M. Zaharia

Complementarias

- Learning Spark, H. Karau, A. Konwinski, P. Wendell, M. Zaharia
- Learning PySpark, T. Drabas, D. Lee
- PySpark Cookbook, D. Lee, T. Drabas