23-24

MÁSTER UNIVERSITARIO EN INDUSTRIA CONECTADA

GUÍA DE ESTUDIO PÚBLICA

PLATAFORMAS PARA PROCESAMIENTO DE DATOS MASIVOS

CÓDIGO 28070060

UNED

23-24

PLATAFORMAS PARA PROCESAMIENTO DE DATOS MASIVOS

CÓDIGO 28070060

ÍNDICE

Nombre de la asignatura	PLATAFORMAS PARA PROCESAMIENTO DE DATOS MASIVOS
Código	28070060
Curso académico	2023/2024
Título en que se imparte	MÁSTER UNIVERSITARIO EN INDUSTRIA CONECTADA
Tipo	CONTENIDOS
Nº ETCS	5
Horas	125.0
Periodo	SEMESTRE 1
Idiomas en que se imparte	CASTELLANO

PRESENTACIÓN Y CONTEXTUALIZACIÓN

PRESENTACIÓN

El trabajo con datos masivos exige la utilización de infraestructuras computacionales específicamente diseñadas para ellos. Estas infraestructuras difieren de las infraestructuras tradicionales en varios aspectos. Para empezar, es necesario combinar la potencia de cómputo de muchos ordenadores, construyendo lo que se conoce como un cluster de ordenadores. Por otro lado, es necesario utilizar paradigmas de programación que puedan aprovechar la potencia de cómputo del cluster pero de una forma sencilla para el desarrollador encargado de implementar los programas para el análisis de datos masivos. Ambos aspectos pueden desarrollarse utilizando servicios de proveedores en la nube. En esta asignatura se muestran algunas de las tecnologías más importantes que permiten desplegar infraestructuras para el procesamiento de datos masivos.

Dentro de este Máster es importante adquirir una visión sólida de las herramientas más utilizadas en ese contexto, dado que son esenciales para mover y tratar datos masivos, tanto estructurados como no estructurados.

CONTEXTUALIZACIÓN

La asignatura de "Plataformas para Procesamiento de Datos Másivos" se trata de una asignatura de 5 créditos ECTS, con carácter optativo, impartida en el primer semestre del Máster Universitario en Industria Conectada. Los estudiantes que cursen esta asignatura optativa adquirirán la siguiente competencia específica "Conocer y ser capaz de usar plataformas para el análisis de datos masivos en contextos de industria conectada".

Esta asignatura guarda relación más directa con las siguientes asignaturas también disponibles en el mismo Máster:

•Computación en la Nube para Entorno Industriales

•Visualización y Analítica de Datos Masivos

REQUISITOS Y/O RECOMENDACIONES PARA CURSAR ESTA ASIGNATURA

Se recomienda que los interesados en cursar el Máster tengan un nivel de lectura en inglés suficiente como para entender contenidos técnicos en dicha lengua. Gran parte de la bibliografía, así como los recursos proporcionados al estudiante en el curso virtual pueden estar únicamente en inglés, debido a la novedad de algunos de los contenidos propuestos

para la asignatura.
Dado que se verán diferentes tipos de despliegues y/o tecnologías, es necesario que los estudiantes dispongan de solidos conocimientos en sistemas operativos y redes, a nivel de comandos de gestión y manipulación de ficheros (especialmente, Linux).
Se fomentará el uso de software libre siempre y cuando sea posible para la realización de las actividades y las practicas propuestas.

EQUIPO DOCENTE

Nombre y Apellidos	RAFAEL PASTOR VARGAS
Correo Electrónico	rpastor@dia.uned.es
Teléfono	91398-8383
Facultad	ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
Departamento	SISTEMAS DE COMUNICACIÓN Y CONTROL

Nombre y Apellidos	RAFAEL PASTOR VARGAS
Correo Electrónico	rpastor@scc.uned.es
Teléfono	91398-8383
Facultad	ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
Departamento	SISTEMAS DE COMUNICACIÓN Y CONTROL

Nombre y Apellidos	ANTONIO ROBLES GOMEZ (Coordinador de asignatura)
Correo Electrónico	arobles@scc.uned.es
Teléfono	91398-8480
Facultad	ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
Departamento	SISTEMAS DE COMUNICACIÓN Y CONTROL

Nombre y Apellidos	AGUSTIN CARLOS CAMINERO HERRAEZ
Correo Electrónico	accaminero@scc.uned.es
Teléfono	91398-9468
Facultad	ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
Departamento	SISTEMAS DE COMUNICACIÓN Y CONTROL

HORARIO DE ATENCIÓN AL ESTUDIANTE

Las consultas sobre los contenidos y funcionamiento de la asignatura se plantearán principalmente en los foros del curso virtual, que serán atendidas por el Equipo Docente de la asignatura.
Para contactar directamente con el Equipo Docente se utilizará preferentemente el correo electrónico, pudiéndose también realizar consultas telefónicas y entrevista personal en los horarios establecidos.
Datos del equipo docente:
Antonio Robles Gómez
Horario: Lunes lectivos de 10 a 14 horas
Email: arobles@scc.uned.es
Tfno: 913988480

Agustín C. Caminero Herráez

Horario: Lunes lectivos de 11 a 13, y de 15 a 17 horas
Email: accaminero@scc.uned.es
Tfno: 91 398 9468

Rafael Pastor Vargas
Horario: Lunes lectivos de 16 a 20 horas
Email: rpastor@scc.uned.es
Tfno: 91 398 8383

También es posible consultar con los docentes en la siguiente dirección postal:
ETSI Informática. UNED.
C/Juan del Rosal 16. 28040. Madrid.

COMPETENCIAS QUE ADQUIERE EL ESTUDIANTE

Competencias Generales (CG):
CG1 - Diseñar estrategias para organizar y planificar entornos industriales conectados
CG2 - Resolver problemas asociados al diseño o desarrollo de sistemas industriales conectados
CG4 - Ser capaz de gestionar información proveniente de sistemas industriales conectados
CG5 - Ser capaz de diseñar y desarrollar sistemas industriales conectados de manera eficiente
Competencias Básicas (CB):
CB6 - Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación
CB7 - Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio
CB8 - Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios
CB9 - Que los estudiantes sepan comunicar sus conclusiones y los conocimientos y razones últimas que las sustentan a públicos especializados y no especializados de un modo claro y sin ambigüedades
CB10 - Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo
Además, los estudiantes que cursen esta asignatura optativa adquirirán la siguiente ***competencia específica***: Conocer y ser capaz de usar plataformas para el análisis de datos masivos en contextos de industria conectada.

RESULTADOS DE APRENDIZAJE

Los resultados que se pretenden alcanzar con el estudio de esta asignatura son los siguientes:

•Distinguir entre las principales herramientas de inyección, programación y almacenamiento de datos masivos, tanto en batch como en streaming. De este modo, se podrán examinar las ventajas y desventajas del uso de un paradigma u otro.

•Diseñar programas para el análisis de datos masivos utilizando las herramientas adecuadas para la inyección, análisis y almacenamiento de dichos datos.

•Describir las características más importantes de las principales arquitecturas de programación de Big Data y de sus formas de despliegue tanto local como en la nube.

•Identificar y seleccionar las diferentes opciones de configuración con el objetivo de optimizar las infraestructuras de Big Data y el desarrollo de algoritmos científicos paralelizables, mejorando así la eficiencia de procesamiento de datos.

CONTENIDOS

Ecosistema Big Data

Se introducirá el ecosistema Big Data de herramientas para el procesamiento paralelo de datos masivo y su programación distribuida. Ventajas y desventajas de las herramientas del ecosistema.

En concreto, estan previstos los siguientes contenidos:

•Introducción a Big Data y Hadoop.

•Programación MapReduce.

•Programación MapReduce con lenguajes de alto nivel: Hive y Pig.

•Herramientas de serialización/deserialización e inyección/extracción de datos.

Técnicas de procesamiento masivo

Se verán técnicas de procesamiento masivo de datos en memoria en tiempo real: componentes y configuración.

En concreto, estan previstos los siguientes contenidos:

•Introducción e instalación de Apache Spark.

•Programación de aplicaciones en Spark.

•Librerías/Componentes de Spark.

•Configuración, monitorización y optimización de Spark.

Gestión de la información en tiempo real

El tema versará sobre la gestión de la información en tiempo real mediante arquitecturas específicas y los eventos generados por las mismas. Análisis de los resultados de salida posibles.
En concreto, estan previstos los siguientes contenidos:
•Introducción a las arquitecturas de procesamiento de streams: Lambda y Kappa.
•Componentes tecnológicos de adquisición y transmisión/distribución de eventos:Kafka.
•Procesamiento de Streams: Spark Streaming.

Servicios en la nube para el almacenamiento y procesamiento de datos

El tema tratará sobre servicios en la nube para el almacenamiento y procesamiento paralelo de datos masivos.
También se explicarán ejemplos de algoritmos paralelizables en entornos industriales: desarrollo de optimizaciones para la obtención de conclusiones.

METODOLOGÍA

Esta asignatura ha sido diseñada para la enseñanza a distancia. Por tanto, el sistema de enseñanza-aprendizaje estará basado en gran parte en el estudio independiente o autónomo del estudiante. Para ello, el estudiante contará con diversos materiales que permitirán su trabajo autónomo y la Guía de Estudio de la asignatura, que incluye orientaciones para la realización de las actividades prácticas. Asimismo, mediante la plataforma virtual de la UNED existirá un contacto continuo entre el equipo docente y los/as estudiantes, así como una interrelación entre los propios estudiantes a través de los foros, importantísimo en la enseñanza no presencial.
El estudio de esta asignatura se realizará a través de los materiales que el Equipo Docente publicará en el curso virtual.
Las actividades formativas para el estudio de la asignatura son las siguientes:
•Estudios de contenidos (50 horas).
•Tutorías en línea (12 horas).
•Actividades en la plataforma virtual: participación en foros y trabajo en grupo (4 horas).
•Activades prácticas / Trabajos:
•Preparación de trabajos a distancia y pruebas de evaluación continua (9 horas).
•Actividades prácticas con simuladores, laboratorios virtuales o remotos (50 horas).
Total: 125 horas
Los medios necesarios para el aprendizaje son:

1. **Materiales teórico-prácticos** preparados por el Equipo Docente para cubrir los conceptos básicos del temario.

2. **Bibliografía complementaria**. El estudiante puede encontrar en ella información adicional para completar su formación.
3. **Curso Virtual de la asignatura**, donde el estudiante encontrará:

•Una guía de la asignatura en la que se hace una descripción detallada del plan de trabajo propuesto.

•Un calendario con la distribución temporal de los temas propuesta por el Equipo Docente y con las fechas de entrega de las actividades teórico-prácticas que el estudiante tiene que realizar para su evaluación.

•Enunciado de las actividades teórico-prácticas propuestas y zona donde depositar los entregables asociados a dichas actividades.

•Los foros por medio de los cuales el Equipo Docente aclarará las dudas de carácter general y que se usarán también para comunicar todas aquellas novedades que surjan a lo largo del curso. Éste será el principal medio de comunicación entre los distintos participantes en la asignatura.

SISTEMA DE EVALUACIÓN

TIPO DE PRUEBA PRESENCIAL

Tipo de examen: No hay prueba presencial

CARACTERÍSTICAS DE LA PRUEBA PRESENCIAL Y/O LOS TRABAJOS

Requiere Presencialidad: No

Descripción

El estudiante debe realizar dos *Actividades prácticas* y un *Trabajo*. Abarcan el 100% de la nota final. En concreto, los elementos de evaluación serán los siguientes:

Memoria de actividad práctica 1: El estudiante deberá realizar un desarrollo consistente en el procesamiento de un dataset utilizando tanto MapReduce como una herramienta de programación de alto nivel. Todo el desarrollo debe estar documentado en una memoria.

Memoria de actividad práctica 2: El estudiante deberá realizar un desarrollo analítico usando Spark y usar alguna de las librerías explicadas en el módulo: Graphx o MLLib (aprendizaje máquina). odo el desarrollo debe estar documentado en una memoria.

Trabajo: El estudiante deberá realizar un trabajo e investigación sobre alguna temática específica de la asignatura, teniendo en cuenta además algún proveedor de la nube para el procesamiento de datos masivos en la nube, y otras tareas relacionadas.

Criterios de evaluación

El equipo docente publicará una guía para su realización, especificando los criterios de evaluación. Es obligatorio realizar y entregar las dos actividades prácticas y el trabajo. Cada elemento se evaluará sobre 10 puntos y es necesario obtener una calificación media mínima de 4 sobre 10 para poder superar la asignatura.

Ponderación de la prueba presencial y/o los trabajos en la nota final	Memoria de actividad práctica 1: 35%; Memoria de actividad práctica 2: 35%; Trabajo: 30%.
Fecha aproximada de entrega	Memoria de actividad práctica 1: noviembre; Memoria de actividad práctica 2: diciembre/enero; Trabajo: enero/febrero.

Comentarios y observaciones

Se podrán entregar además en la convocatoria extraordinaria, con la fecha que indique el equipo docente.

PRUEBAS DE EVALUACIÓN CONTINUA (PEC)

¿Hay PEC? No

Descripción

Criterios de evaluación

Ponderación de la PEC en la nota final

Fecha aproximada de entrega

Comentarios y observaciones

OTRAS ACTIVIDADES EVALUABLES

¿Hay otra/s actividad/es evaluable/s? No

Descripción

Criterios de evaluación

Ponderación en la nota final

Fecha aproximada de entrega

Comentarios y observaciones

¿CÓMO SE OBTIENE LA NOTA FINAL?

La asignatura se evaluará en base a los siguientes elementos:
Memoria de la actividad práctica 1: Valdrá un 35 % de la nota final.
Memoria de la actividad práctica 2: Valdrá un 35 % de la nota final.
Trabajo: Valdrá un 30 % de la nota final.

Se deben tener en cuenta las siguientes observaciones:

Si en cada elemento de evaluación (***desarrollado de manera individual***) no se obtiene al menos el 40% de la puntuación individual total de cada una de ellas, entonces el/la estudiante estará suspenso.

En otro caso (se tiene más del 40% de la nota total para cada uno de los elementos de evaluación, todos son obligatorios), se calculará la nota final sumando las diferentes calificaciones ponderadas con los porcentajes descritos más arriba.

Aprobarán la asignatura los estudiantes que consigan al menos 5 puntos en la nota final calculada con las ponderaciones definidas más arriba.

BIBLIOGRAFÍA BÁSICA

La bibliografía básica será proporcionada al estudiante dentro del curso virtual, estará compuesta por materiales teórico-prácticos propuestos por el equipo docente.
Gran parte de la bibliografía, así como los recursos proporcionados al estudiante en el curso virtual pueden estar únicamente en inglés, debido a la actualidad de los contenidos propuestos para la asignatura.

BIBLIOGRAFÍA COMPLEMENTARIA

El Equipo Docente propone una serie de libros disponibles de forma gratuíta dentro de la biblioteca digital de la UNED. Se proporcionan enlaces a los libros que funcionan tras autenticarse en UNED.es:

Título: MapReduce Design Patterns
Autores: Donald Miner; Adam Shook
Editorial: O'Reilly Media, Inc.
Año, 2012
ISBN-13 en papel: 978-1-4493-2717-0
URL (solamente funciona tras autenticarse en UNED.es):
https://learning.oreilly.com/library/view/mapreduce-design-patterns/9781449341954/

Título: Hadoop: The Definitive Guide, 4th Edition
Autor:Tom White
Editorial: O'Reilly Media, Inc.
Año: 2015
ISBN-13 en papel: 978-1-4919-0163-2
URL (solamente funciona tras autenticarse en UNED.es):
https://learning.oreilly.com/library/view/hadoop-the-definitive/9781491901687/

Título: Practical Hadoop Ecosystem: A Definitive Guide to Hadoop-Related Frameworks and Tools
Autor: Deepak Vohra
Editorial: Apress
Año: 2016
ISBN-13 en papel: 978-1-4842-2198-3
URL (solamente funciona tras autenticarse en UNED.es):
https://learning.oreilly.com/library/view/practical-hadoop-ecosystem/9781484221990/

Título: Designing Data-Intensive Applications
Autor: Martin Kleppmann
Editorial: O'Reilly Media, Inc.
Año: 2017
ISBN-13 en papel: 978-1-4493-7332-0
URL (solamente funciona tras autenticarse en UNED.es):
https://learning.oreilly.com/library/view/designing-data-intensive-applications/9781491903063/

Titulo: Apache Hive Cookbook
Autores: Hanish Bansal; Saurabh Chauhan; Shrey Mehrotra
Editorial: Packt Publishing
Año: 2016.
ISBN-13 en papel: 978-1-78216-108-0
ISBN-13 web: 978-1-78216-109-7
URL (solamente funciona tras autenticarse en UNED.es):
https://learning.oreilly.com/library/view/apache-hive-cookbook/9781782161080/

Título: Hadoop with Python
Autores: Zachary Radtka, Donald Miner
Editorial: O’Reilly
Año: 2015
ISBN: 978-1-491-94227-7
URL (solamente funciona tras autenticarse en UNED.es):
https://learning.oreilly.com/library/view/hadoop-with-python/9781492048435/

Título: Fast Data Processing with Spark 2 -Third Edition
Autor: Krishna SankarEditorial: Packt Publishing
Año:2016ISBN-13 en papel:978-1-78588-927-1
URL (solamente funciona tras autenticarse en UNED.es):
https://learning.oreilly.com/library/view/fast-data-processing/9781785889271/

Título: Sams Teach Yourself Apache Spark™ in 24 Hours

Autor: Jeffrey Aven.
Editorial: SamsAño:2016.
ISBN-13 en papel:978-0-672-33851-9.
URL (solamente funciona tras autenticarse en UNED.es):
https://learning.oreilly.com/library/view/sams-teach-yourself/9780134445786/

Título: Mastering Apache Spark 2.x -Second Edition
Autor: Romeo Kienzler
Editorial: Packt Publishing
Año:2017
ISBN-13 en papel:978-1-78646-274-9
URL (solamente funciona tras autenticarse en UNED.es):
https://learning.oreilly.com/library/view/mastering-apache-spark/9781786462749/

Título: Apache Spark 2.x Cookbook
Autor: Rishi Yadav
Editorial: Packt Publishing
Año: 2017
ISBN-13 en papel:978-1-78712-726-5
URL (solamente funciona tras autenticarse en UNED.es):
https://learning.oreilly.com/library/view/apache-spark-2x/9781787127265/

Título: Spark for Python Developers
Autor: Amit Nandi
Editorial: Packt Publishing
Año: 2015
ISBN-13 en Web: 978-1-78439-737-1
URL (solamente funciona tras autenticarse en UNED.es):
https://learning.oreilly.com/library/view/spark-for-python/9781784399696/

Título: Machine Learning with Spark -Second Edition
Autor:Rajdeep Dua; Manpreet Singh Ghotra; Nick Pentreath
Editorial: Packt Publishing
Año: 2017
ISBN-13 en papel:978-1-78588-993-6
URL (solamente funciona tras autenticarse en UNED.es):
https://learning.oreilly.com/library/view/machine-learning-with/9781785889936/

Título: Spark GraphX in Action
Autor:Michael S. Malak and Robin East
Editorial: Manning Publications
Año:2016
ISBN-13: 978-1-61729-252-1
URL (solamente funciona tras autenticarse en UNED.es):
https://learning.oreilly.com/library/view/spark-graphx-in/9781617292521/

Título: Spark in Action
Autor:Petar Zeevi, Marko Bonai
Editorial: Manning Publications
Año: 2016
ISBN-13: 978-1-61729-260-6
URL (solamente funciona tras autenticarse en UNED.es):
https://learning.oreilly.com/library/view/spark-in-action/9781617292606/

Título: Streaming Systems
Autor: Reuven Lax, Slava Chernyak, Tyler Akidau
Editorial: O'Reilly Media, Inc.
Año: 2018
ISBN-13: 978-1-49198-387-4
URL (solamente funciona tras autenticarse en UNED.es):
https://learning.oreilly.com/library/view/streaming-systems/9781491983867/

Título: Kafka: The Definitive Guide
Autor: Gwen Shapira, Neha Narkhede, Todd Palino
Editorial: O'Reilly Media, Inc.
Año: 2017
ISBN-13: 978-1-49193-616-0
URL (solamente funciona tras autenticarse en UNED.es):
https://learning.oreilly.com/library/view/kafka-the-definitive/9781491936153/

Título: Stream Processing with Apache Spark
Autor: Francois Garillot, Gerard Maas
Editorial: O'Reilly Media, Inc.
Año: 2019

ISBN-13: 978-1-49194-424-0
URL (solamente funciona tras autenticarse en UNED.es):
https://learning.oreilly.com/library/view/stream-processing-with/9781491944233/

RECURSOS DE APOYO Y WEBGRAFÍA

Los/as estudiantes dispondrán de los siguientes recursos de apoyo al estudio:

- **Guía de la asignatura.** Incluye el plan de trabajo y orientaciones para su desarrollo. Esta guía será accesible desde el curso virtual.
- **Curso virtual.** A través de esta plataforma los/as estudiantes tienen la posibilidad de consultar información de la asignatura, realizar consultas al Equipo Docente a través de los foros correspondientes, consultar e intercambiar información con el resto de los compañeros/as.
- **Biblioteca.** El estudiante tendrá acceso tanto a las bibliotecas de los Centros Asociados como a la biblioteca de la Sede Central, en ellas podrá encontrar un entorno adecuado para el estudio, así como de distinta bibliografía que podrá serle de utilidad durante el proceso de aprendizaje. Además, desde la biblioteca digital de la UNED, el estudiante tendrá acceso a Safari Books Online, una biblioteca digital con más de 30.000 libros técnicos en constante actualización.

PRÁCTICAS DE LABORATORIO

¿Hay prácticas en esta asignatura de cualquier tipo (en el Centro Asociado de la Uned, en la Sede Central, Remotas, Online,..)?

Si, Online/Remotas

CARACTERÍSTICAS GENERALES

Presencial: No

Obligatoria: Si

Es necesario aprobar el examen para realizarlas: No, no hay examen presencial.

Fechas aproximadas de realización: Durante el periodo lectivo (especificado en Sistema de evaluación).

Se guarda la nota en cursos posteriores si no se aprueba el examen:
(Si es así, durante cuántos cursos): No

Cómo se determina la nota de las prácticas: Especificado en "Sistema de evaluación".

REALIZACIÓN

Lugar de realización (Centro Asociado/ Sede central/ Remotas/ Online): Online

N.º de sesiones: Tres entregas.

Actividades a realizar: Especificado con detalle en "Sistema de evaluación".

IGUALDAD DE GÉNERO

En coherencia con el valor asumido de la igualdad de género, todas las denominaciones que en esta Guía hacen referencia a órganos de gobierno unipersonales, de representación, o miembros de la comunidad universitaria y se efectúan en género masculino, cuando no se hayan sustituido por términos genéricos, se entenderán hechas indistintamente en género femenino o masculino, según el sexo del titular que los desempeñe.