



Guía docente				
Datos Identificativos				2018/19
Asignatura (*)	Análisis de Datos con HPC		Código	614473108
Titulación	Mestrado Universitario en Computación de Altas Prestacións / High Performance Computing (Mod. Presencial 2018)			
Descriptor				
Ciclo	Periodo	Curso	Tipo	Créditos
Máster Oficial	2º cuatrimestre	Primero	Optativa	6
Idioma	Inglés			
Modalidad docente	Presencial			
Prerrequisitos				
Departamento	Enxeñaría de Computadores			
Coordinador/a	López Taboada, Guillermo	Correo electrónico	guillermo.lopez.taboada@udc.es	
Profesorado	López Taboada, Guillermo	Correo electrónico	guillermo.lopez.taboada@udc.es	
	Rodríguez Álvarez, Gabriel		gabriel.rodriguez@udc.es	
Web	aula.cesga.es			
Descripción general	La cantidad cada vez mayor de información accesible a través de Internet hace que el procesamiento eficiente de grandes cantidades de datos sea cada vez de mayor interés. Esto ha llevado al desarrollo de nuevas técnicas de almacenamiento y procesamiento de ingentes cantidades de información, denominadas técnicas Big Data, que se adaptan de forma natural a los sistemas distribuidos.			
	El objetivo principal de esta materia es dar a conocer diferentes técnicas de procesamiento de grandes cantidades de información dentro del mundo Big Data, en particular en el ámbito del ecosistema Hadoop, y hacer una comparación con el tipo de procesamiento más tradicional del mundo HPC para, desde una actitud reflexiva, poder seleccionar las herramientas óptimas para resolver un determinado problema.			

Competencias del título	
Código	Competencias del título
A1	CE1 - Definir, evaluar y seleccionar la arquitectura y el software más adecuado para la resolución de un problema
A2	CE2 - Analizar y mejorar el rendimiento de una arquitectura o un software dado
B1	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
B2	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
B6	CG1 - Ser capaz de buscar y seleccionar la información útil necesaria para resolver problemas complejos, manejando con soltura las fuentes bibliográficas del campo
B8	CG3 - Ser capaz de mantener y extender planteamientos teóricos fundados para permitir la introducción y explotación de tecnologías nuevas y avanzadas en el campo
B10	CG5 - Ser capaz de trabajar en equipo, especialmente de carácter multidisciplinar, y ser hábiles en la gestión del tiempo, personas y toma de decisiones.
C1	CT1 - Utilizar las herramientas básicas de las tecnologías de la información y las comunicaciones (TIC) necesarias para el ejercicio de su profesión y para el aprendizaje a lo largo de su vida
C4	CT4 - Valorar la importancia que tiene la investigación, la innovación y el desarrollo tecnológico en el avance socioeconómico y cultural de la sociedad

Resultados de aprendizaje	
Resultados de aprendizaje	Competencias del título



El alumno será capaz de seleccionar, instalar, configurar y gestionar el software básico para el procesamiento de datos masivos.	AP1 AP2	BP2 BP6 BP8 BP10	CP1
El alumno será capaz de implementar códigos en algún lenguaje especializado en el procesamiento de datos masivos.	AP2	BP1 BP2 BP10	CP1
El alumno conocerá y aprenderá a utilizar algunas de las herramientas disponibles para Data Engineering (en particular, par Ingesta/Almacenamiento/Procesado/Visualización).	AP1 AP2	BP1 BP2	CP1 CP4
El alumno adquirirá la habilidad necesaria para la búsqueda, selección y manejo de recursos (bibliografía, software, etc.) relacionados con Big Data.	AP1 AP2	BP1 BP6	CP1 CP4

Contenidos	
Tema	Subtema
1. Introducción a Data Engineering	1.1 HPC vs Big Data: similitudes y diferencias en el tratamiento de datos 1.2 Tecnologías Hardware y Software para High Performance Data Engineering 1.3 Data Engineering en infraestructuras HPC vs entornos Cloud
2 Etapas de Data Engineering	2.1 Modelado (Formatos, Compresión, Diseño de Esquemas) 2.2 Ingesta (Periodicidad, Transformaciones, Herramientas) 2.3 Almacenamiento (HDFS y BBDD NoSQL, HBase, MongoDB, Cassandra) 2.4 Procesado (Batch, Real-Time) 2.5 Orquestación 2.6 Análisis (SQL, Machine Learning, Graphs, UI) 2.7 Gobernanza 2.8 Integración con BI (Visualización)
3 Introducción a Analítica de Datos	3.1 Exploratory Data Analytics 3.2 Introducción a Machine Learning
4 Casos de Uso	4.1 Aplicaciones en Internet de las Cosas (entornos Smart e Industria 4.0) 4.2 Aplicaciones en ciencias e ingeniería

Planificación				
Metodologías / pruebas	Competencias	Horas presenciales	Horas no presenciales / trabajo autónomo	Horas totales
Sesión magistral	A2 A1 B1 C4	18	0	18
Prácticas de laboratorio	B1 B8 B10	20	60	80
Trabajos tutelados	A1 A2 B1 B2 B8	0	45	45
Discusión dirigida	B6 C1 C4	4	3	7
Atención personalizada		0		0
(*)Los datos que aparecen en la tabla de planificación són de carácter orientativo, considerando la heterogeneidad de los alumnos				

Metodologías	
Metodologías	Descripción
Sesión magistral	Impartidas por los profesores de la asignatura. Incluyen exposición de material teórico, así como de seminarios.
Prácticas de laboratorio	Resolución de problemas y casos prácticos.
Trabajos tutelados	Realización de prácticas de mayor entidad de forma semiautónoma, guiados por los profesores de la asignatura.
Discusión dirigida	Orientación para la realización de los trabajos individuales o en grupo, resolución de dudas y actividades de evaluación continua.



Atención personalizada

Metodologías	Descripción
Discusión dirigida Prácticas de laboratorio Trabajos tutelados	Durante las prácticas de laboratorio, trabajos tutelados, y discusiones dirigidas, los estudiantes podrán presentar preguntas, dudas, etc. El profesor, atendiendo a sus solicitudes, repasará conceptos, resolverá nuevos problemas o utilizará cualquier actividad que considere adecuada para resolver las cuestiones planteadas.

Evaluación

Metodologías	Competencias	Descripción	Calificación
Discusión dirigida	B6 C1 C4	Seguimiento continuo y objetivable de una participación activa.	10
Prácticas de laboratorio	B1 B8 B10	Evaluación de las prácticas llevadas a cabo por los estudiantes.	40
Trabajos tutelados	A1 A2 B1 B2 B8	Evaluación de los trabajos tutelados desarrollados por los estudiantes.	50

Observaciones evaluación

Primera oportunidad (ordinaria - mayo):

Realización de prácticas: 40% Trabajos académicamente dirigidos: 50% Seguimiento continuo y objetivable de una participación activa: 10% Segunda oportunidad (extraordinaria - julio):

Realización de prácticas: el alumno conservará la nota obtenida en este apartado en la convocatoria ordinaria, ya que para la convocatoria extraordinaria no se planifican nuevas actividades. 40% de la nota final. Evaluación de trabajos académicamente dirigidos: será preciso presentar los trabajos académicamente dirigidos que no se hubieran presentado en la convocatoria ordinaria y se volverán a presentar, tras las modificaciones oportunas indicadas por los profesores, aquellos que no hubiesen recibido una calificación mínima necesaria para aprobar en dicha convocatoria. 50% de la nota final. Seguimiento continuo y objetivable de una participación activa: el alumno conservará la nota obtenida en este apartado en la convocatoria ordinaria, ya que para la convocatoria extraordinaria no se planifican nuevas actividades. 10% de la nota final. No presentado:

Se considerará no presentado al alumno que no entregue ninguna práctica ni trabajo académicamente dirigido.

Fuentes de información

Básica	- Tom White (2015). Hadoop: The Definitive Guide. O'Reilly (4ª ed.) - Wes McKinney (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly (2ª ed.)
Complementaria	- Alex Holmes (2014). Hadoop in practice. Manning (2ª ed.)

Recomendaciones

Asignaturas que se recomienda haber cursado previamente

Asignaturas que se recomienda cursar simultáneamente

Asignaturas que continúan el temario

Otros comentarios

Recomendaciones para el estudio de la materia Debido al fuerte componente práctico es recomendable ir haciendo las actividades prácticas y trabajos académicamente dirigidos de forma regular a lo largo del cuatrimestre. El conocimiento del inglés tanto hablado como escrito es imprescindible dado que la bibliografía y las conferencias externas pueden desarrollarse en inglés. Observaciones Se hará un uso intensivo de herramientas de comunicación online: videoconferencia, chat, etc. Las sesiones presenciales serán grabadas para u revisión posterior. Además, se hará uso de la herramienta Aula CESGA para la distribución de contenidos, creación de foros de discusión, etc... Las herramientas software utilizadas en esta materia son de fuentes abiertas.



(*) La Guía Docente es el documento donde se visualiza la propuesta académica de la UDC. Este documento es público y no se puede modificar, salvo cosas excepcionales bajo la revisión del órgano competente de acuerdo a la normativa vigente que establece el proceso de elaboración de guías