Guía Docente				
Datos Identificativos				2022/23
Asignatura (*)	Análise de Datos con HPC		Código	614973108
Titulación	Mestrado Universitario en Computación de Altas Prestacións / High Performance Computing (Mod. Virtual)			
Descriptores				
Ciclo	Período	Curso	Tipo	Créditos
Mestrado Oficial	2º cuadrimestre	Primeiro	Optativa	6
Idioma	Inglés			
Modalidade docente	Non presencial			
Prerrequisitos				
Departamento	Departamento profesorado másterEnxeñaría de Computadores			
Coordinación	López Taboada, Guillermo	Correo electrónico	guillermo.lopez.taboada@udc.es	
Profesorado	López Taboada, Guillermo Rodríguez Álvarez, Gabriel	Correo electrónico	guillermo.lopez.taboada@udc.es gabriel.rodriguez@udc.es	
Web	aula.cesga.es			
Descrición xeral	A cantidade cada vez maior de información accesible a través de Internet fai que o procesamento eficiente de grandes cantidades de datos sexa cada vez de maior interese. Isto levou ao desenvolvemento de novas técnicas de almacenamento e procesamento de inxentes cantidades de información, denominadas técnicas Big Data, que se adaptan de forma natural aos sistemas distribuídos.			

Competencias do título	
Código	Competencias do título
A1	CE1 - Definir, avaliar e seleccionar a arquitectura e o software máis axeitado para a resolución dun problema
A2	CE2 - Analizar e mellorar o rendimento dunha arquitectura ou un software dado
B1	CB6 - Posuir e comprender coñecementos que aporten unha base ou oportunidade de ser orixinais no desenrolo e/ou aplicación de ideas, a miudo nun contexto de investigación
B2	CB7 - Que os estudantes saiban aplicar os coñecementos adquiridos e súa capacidade de resolución de problemas en contornas novas ou pouco coñecidas dentro de contextos máis amplos (ou multidisciplinares) relacionados coa súa área de estudo
B6	CG1 - Ser capaz de buscar e seleccionar a información útil necesaria para resolver problemas complexos, manexando con soltura as fontes bibliográficas do campo
B8	CG3 - Ser capaz de manter e extender plantexamentos teóricos fundados para permitir a introducción e explotación de tecnoloxías novas e avanzadas no campo
B10	CG5 - Ser capaz de traballar en equipo, especialmente de carácter multidisciplinar, e ser hábiles na xestión do tempo, persoas e toma de decisións.
C1	CT1 -Utilizar as ferramentas básicas das tecnoloxías da información e as comunicacións (TIC) necesarias para o exercicio da súa profesión e para a aprendizaxe ao longo da súa vida.
C4	CT4 - Valorar a importancia que ten a investigación, a innovación e o desenrolo tecnolóxico no avance socioeconómico e cultural da sociedade

Resultados da aprendizaxe			
Resultados de aprendizaxe	Competencias do título		
O alumno será capaz de seleccionar, instalar, configurar e xestionar o software básico para o procesamento de datos masivos.	AP1 AP2	BP2 BP6 BP8 BP10	CP1
O alumno será capaz de implementar códigos nalgunha linguaxe especializada no procesamento de datos masivos.	AP2	BP1 BP2 BP10	CP1

O alumno coñecerá e aprenderá a utilizar algunhas das ferramentas dispoñibles para Data Engineering (en particular, para Inxesta/Almacenamento/Procesado/Visualización).	AP1 AP2	BP1 BP2	CP1 CP4
O alumno adquirirá a habilidade necesaria para a procura, selección e manexo de recursos (bibliografía, software, etc.) relacionados con Big Data.	AP1 AP2	BP1 BP6	CP1 CP4

Contidos	
Temas	**Subtemas**
1. Introducción a Data Engineering	1.1 HPC vs Big Data: similitudes e diferencias no tratamento de datos 1.2 Tecnoloxías Hardware e Software para High Performance Data Engineering 1.3 Data Engineering en infraestructuras HPC vs entornos Cloud
2. Introducción a Analítica de Datos	2.1 Exploratory Data Analytics 2.2 Introducción a Machine Learning
3. Etapas de Data Engineering	3.1 Modelado (Formatos, Compresión, Deseño de Esquemas) 3.2 Inxesta (Periodicidade, Transformaciones, Ferramentas) 3.3 Almacenamento (HDFS y BBDD NoSQL, HBase, MongoDB, Cassandra) 3.4 Procesado (Batch, Real-Time) 3.5 Orquestación 3.6 Análise (SQL, Machine Learning, Graphs, UI) 3.7 Gobernanza 3.8 Integración con BI (Visualización)
4. Casos de Uso	4.1 Aplicacións en Internet das Cosas (entornos Smart e Industria 4.0) 4.2 Aplicacións en ciencias e enxeñaría

Planificación				
Metodoloxías / probas	**Competencias**	**Horas presenciais**	**Horas non presenciais / traballo autónomo**	**Horas totais**
Lecturas	A1 A2 B1 B6 C4	0	18	18
Prácticas de laboratorio	B1 B8 B10	0	80	80
Traballos tutelados	A1 A2 B1 B2 B8	0	45	45
Discusión dirixida	B6 C1 C4	4	2	6
Atención personalizada		1	0	1
***Os datos que aparecen na táboa de planificación son de carácter orientativo, considerando a heteroxeneidade do alumnado**				

Metodoloxías	
Metodoloxías	**Descrición**
Lecturas	Instrucción programada a través de materiais docentes.
Prácticas de laboratorio	Resolución de problemas e casos prácticos.
Traballos tutelados	Realización de prácticas de mayor entidade de forma semiautónoma, guiados polos profesores da asignatura.
Discusión dirixida	Orientación para a realización dos traballos individuais ou en grupo, resolución de dúbidas e actividades de avaliación continua.

Atención personalizada	
Metodoloxías	**Descrición**
Prácticas de laboratorio Traballos tutelados Discusión dirixida	Durante as prácticas de laboratorio, traballos tutelados, e discusións dirixidas, os estudantes poderán presentar preguntas, dúbidas, etc. O profesor, atendendo ás súas solicitudes, repasará conceptos, resolverá novos problemas ou utilizará calquera actividade que considere adecuada para resolver as cuestions expostas.

Avaliación			
Metodoloxías	Competencias	Descrición	**Cualificación**
Prácticas de laboratorio	B1 B8 B10	Evaluación de las prácticas llevadas a cabo por los estudiantes.	50
Traballos tutelados	A1 A2 B1 B2 B8	Evaluación de los trabajos tutelados desarrollados por los estudiantes.	50

Observacións avaliación
Non presentado: Considerarase non presentado @ alumn@ que non entregue ningunha práctica nin traballo academicamente dirixido. Segunda oportunidade (extraordinaria - xuño/xullo): Volver a realizar aquelas prácticas e traballos tutelados que non se entregaran ou versións melloradas dos xa entregados. Para os casos de realización fraudulenta de exercicios ou probas será de aplicación o recollido na Normativa de avaliación do rendemento académico dos estudantes e de revisión de cualificacións.

Fontes de información	
Bibliografía 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.)
Bibliografía complementaria	- Alex Holmes (2014). Hadoop in practice. Manning (2ª ed.)

Recomendacións
Materias que se recomenda ter cursado previamente
Materias que se recomenda cursar simultaneamente
Materias que continúan o temario
Observacións
Recomendacións para o estudo da materiaDebido ao forte compoñente práctico é recomendable ir facendo as actividades prácticas e traballos academicamente dirixidos de forma regular ao longo do cuadrimestre.O coñecemento do inglés tanto falado como escrito é imprescindible dado que a bibliografía e as conferencias externas poden desenvolverse en inglés.ObservacionsFarase un uso intensivo de ferramentas de comunicación online: videoconferencia, chat, etc. As sesións presenciais serán gravadas para ou revisión posterior. Ademais, farase uso da ferramenta Aula CESGA para a distribución de contidos, creación de foros de discusión, etc...As ferramentas software utilizadas nesta materia son xeralmente open-source ou teñen licencia gratuita para estudantes.