MASTER	Master en Data Science para Finanzas
ASIGNATURA	***Extracción, transformación y carga - ETL***
Nº de ECTS	**5**
Nº de horas docentes	**50**
Nº de horas actividades académicas dirigidas	
Profesores responsables de la asignatura	**Oscar Sánchez y Gabriel Valverde**
Curso académico	**2016 / 2017**
Cuatrimestre	**1º Cuatrimestre**

1.- DESCRIPCIÓN GENERAL DE LA ASIGNATURA Y OBJETIVOS DE DOCENCIA:

En esta asignatura se pretende introducir al alumnado herramientas Big Data de procesamiento en paralelo (Hadoop y Spark) y su aplicación a los sistemas de información en el procesamiento de grandes volúmenes de datos.

Objetivos: Adquisición de conocimientos sobre tecnologías Big Data que permitan al alumno realizar análisis y gestión de datasets de gran tamaño y de diversa naturaleza.

Requisitos: programación con lenguajes R y Python y conocimiento de Linux.

2.- FORMA DE EVALUACIÓN PREVISTA:

Participación y asistencia	10%
Actividades académicas dirigidas	40%
Prueba objetiva final (Trabajo final con presentación)	50%

PROGRAMA DETALLADO		
Nº de sesión	**Detalle del contenido docente: temas, casos prácticos, actividades académicas dirigidas que se verán en dicha sesión,…**	**Lecturas recomendadas o referencias bibliográficas relativas a los conceptos-temas desarrollados en la sesión**
1	Introducción al ecosistema Big Data	Big Data For Dummies (by Judith Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman)
2	Introducción a la herramienta Hadoop	Hadoop For Dummies (by Dirk deRoos)
3	Procesamiento de datos y ETL SingleNode	
4	Entorno R para configurar Hadoop SingleNode	Big Data Analytics with R and Hadoop (by Vignesh Prajapati)

5	Configurar Hadoop SingleNode en R	
6	Entorno Python para configurar Hadoop SingleNode	Hadoop with Python (by Donald Miner)
7	Configurar Hadoop SingleNode en Python	
8	Introducción al entorno Cloud Computing	Amazon Web Services For Dummies (by Bernard Golden)
9	Servicios ofrecidos por AWS, Google Cloud, …	
10	Configuración entorno Cloud	
11	Procesamiento de datos y ETL MultiNode	
12	Entorno R para configurar Hadoop MultiNode	
13	Configurar Hadoop MultiNode en R	
14	Entorno Python para configurar Hadoop MultiNode	
15	Configurar Hadoop MultiNode en Python	
16	Introducción a ETL , procesamiento en paralelo y MapReduce. ELT vs ETL en BigData	
17	Introducción a Spark. ¿Qué es?, Hadoop vs Spark, Carácteristicas de Spark, Spark Ecosystem	How companies are using Spark (Ben Lorica)
18	Introducción a Spark conexión con fuentes de datos	Spark In Action (Petar Zečević and Marko Bonaći)
19	Introducción a Spark en R y Python	
20	Spark SQL	
21	Spark SQL en R y Python	
22	Spark Streaming	Fast Data Processing with Spark (Holden Karau) In-Stream Big Data Processing (Llya Katsov)
23	Spark Streaming en R y Python	Spark Workshop Hosted by Stanford ICME (Reza Zadeh, Matei Zaharia, Ion Stoica)

24	Spark Machine Learning	Machine Learning with Spark (Nick Pentreath) Spark GraphX in Action (Michael S. Malak and Robin East)
25	Spark Machine Learning en R y Python	

INFORMACION ADICIONAL	
Bibliografía básica	Big Data Analytics with R and Hadoop (by Vignesh Prajapati); Spark In Action (Petar Zečević and Marko Bonaći)
Bibliografía Complementaria	La indicada en el programa
Actividades Complementarias	Prácticas en clase
Localización del profesor	gvalverd@ucm.es, oscsanch@ucm.es