UNIVERSIDAD COMPLUTENSE MADRID

Curso de Formación Continua UCM

Big Data y Análisis de Datos con Hadoop y Spark

Código: 0588-C588
Dirección: Luis Llana Díaz y Carlos Gregorio Rodríguez
Duración: 25 horas presenciales
Precio: 220€
Plazas: 21
Matrícula: `https://ucm.es/estudiantes-fc`

Motivación En los últimos años ha crecido enormemente la demanda de profesionales en las áreas de Big Data y Data Science. Las ofertas de empleo reflejan el interés que tienen empresas e instituciones en encontrar especialistas en el uso de tecnologías basadas en clusters de computadores.

Sin embargo, por la relativa novedad , los estudiantes de la Universidad Complutense tienen pocas oportunidades de formarse en dichas herramientas y tecnologías en los planes de estudio actuales.

Este curso ofrece un complemento perfecto a estudiantes de Ciencias e Ingenierías de nuestra universidad para adquirir los conceptos esenciales y conocer las herramientas de programación más utilizadas y demandadas en la actualidad para el trabajo con datos: los clusters basados en Hadoop y la programación paralela con Spark.

Perfil El Big Data suscita mucho interés en áreas de conocimientos muy diversas. Las herramientas y técnicas que exploramos en el curso tienen que ver con la parte de gestión de sistemas de ficheros distribuidos y, sobre todo, con la programación para utilizar la poténcia de cómputo de clústers. El curso es muy adecuado para estudiantes de grado y máster en las ramas de ciencias e ingenierías que hayan cursado alguna asignatura de programación.

Objetivos

- Comprender el marco conceptual y los retos del Big Data
- Entender cómo el uso de clusters y la programación paralela aportan una solución para trabajar con grandes volúmenes de datos
- Conocer Apache Hadoop, Apache Spark y el ecosistema de herramientas asociado
- Entender el diseño del sistema de ficheros distribuido Hadoop HDFS
- Utilización del sistema de ficheros distribuido Hadoop HDFS
- Ajuste de parámetros para mejorar la eficiencia de las tareas en sistemas Hadoop HDFS
- Entender el esquema de programación MapReduce (Hadoop MapReduce)
- Escribir código en Python para solucionar problemas utilizando el esquema MapReduce
- Conocer las características de Apache Spark
- Programar en Python para Apache Spark
- Ajuste de parámetros para mejorar la eficiencia de tareas Spark
- Módulos de Spark para dominios particulares: SQL, MLlib

UNIVERSIDAD COMPLUTENSE MADRID

Curso de Formación Continua UCM

Big Data y Análisis de Datos con Hadoop y Spark

Temario

- Introducción a Python como lenguaje para Data Science
- Introducción al Big Data: conceptos, problemas y soluciones Hw y Sw
- Hadoop Distributed File System (HDFS). Conceptos y utilización
- Esquema de programación Map Reduce sobre clusters HDFS
- Monitorización y ajuste de parámetros de tareas Map Reduce en Hadoop-HDFS
- La abstracción RDD, DataFrame y DataSet en Spark
- Programación en Spark
- Monitorización de tareas de Spark en clusters
- Módulos avanzados de Spark: Spark streaming, Spark MLlib, Spark SQL...

Metodología El curso es eminentemente práctico. Todas las sesiones tendrán lugar en un aula con ordenadores y con todas las herramientas que se estudian instaladas. Después de una introducción a los conceptos esenciales, las participantes trabajarán sobre ejercicios propuestos, casos de estudios y prácticas con las que afianzarán sus habilidades y profundizarán sus conocimientos.

En todo momento contaremos con el apoyo de un campus virtual en el que los participantes tendrán todos los materiales disponibles y que será utilizado también como herramienta de comunicación para resolver dudas y problemas fuera del horario de las sesiones.

Evaluación Puesto que el curso es presencial, con un claro objetivo práctico y que se desarrolla en laboratorio, proponemos una evaluación continua basada en la entrega de los ejercicos, prácticas y trabajos propuestos a las participantes.

Sw/Hw El estudiante no necesita material informático ni hardware ni software pues el curso se desarrollará en laboratorios de la UCM. En todo caso, las herramientas informáticas utilizadas tienen licencia de software libre y por tanto pueden ser instaladas por los alumnos que lo deseen en sus propios ordenadores. Aunque la tecnología con la que trabajamos en el curso se basa en el uso de un clusters de ordenadores, muchas de las herramientas pueden ser usadas o simuladas localmente en una única máquina, lo que posibilita el estudio y aprendizaje de las mismas sin necesitar tantos recursos.

Reconocimiento Los participantes que superen el curso obtendrán un Certificado Académico UCM. Además el curso supone un reconocimiento de 1,5 créditos optativos.

Becas Ofertamos 2 becas totales y 2 becas parciales por curso. Se valorarán el expediente académico y los ingresos. En todo caso, es necesario realizar el pago para formalizar la matrícula. En caso de concesión de beca, este importe es devuelto.