FICHA DE ASIGNATURA

Título: Métodos para extracción, procesamiento y almacenamiento de datos masivos desde Internet

Descripción:

Internet es un canal por el que fluye una cantidad cada vez mayor de información. Ello ha despertado el interés por conseguir un procesamiento eficiente de estas grandes cantidades de datos, mediante técnicas que se adapten a la naturaleza distribuida de la web. El objetivo principal de esta materia es dar a conocer a los alumnos las técnicas de búsqueda, extracción y procesamiento de datos masivos localizados en entornos web. Además, se presenta el framework Hadoop, que proporciona sistemas y técnicas para el almacenamiento y el procesamiento distribuido de grandes cantidades de datos. La asignatura busca familiarizar al estudiante con las principales utilidades integradas dentro del ecosistema Hadoop, como son HDFS (Hadoop Distributed File System), el sistema de archivos que utiliza Hadoop para el almacenamiento de datos; MapReduce, el paradigma de programación ideado por Google en 2004 y empleado por Hadoop para el procesamiento de datos de forma paralela; Hive, un lenguaje similar a SQL para realizar consultas de datos; o Pig, un lenguaje de script para realizar análisis de datos de forma sencilla.

Carácter: Obligatoria

Créditos ECTS: 6

Modalidad: Online

Temario:

1. Fuentes de datos en la nube de Linked Data.
2. Cloud Computing: modelos de servicio y modelos de despliegue.
3. Los proveedores Cloud. Proveedores públicos: Amazon Web Services, Microsoft Azure, Google Compute Engine.
4. Gestores de plataformas Cloud (on-premise): OpenNebula, OpenStack, Eucalyptus.
5. Gestión de computación en la nube: Amazon EC2 y modelos de programación para computación distribuida.
6. La inteligencia de fuentes abiertas (Open Data).
7. Internet de las cosas.
8. Gestión de datos en la nube. Amazon S3, EBS.
9. Sistemas para el procesamiento distribuido de grandes datos con el ecosistema Hadoop. Sus componentes.
10. Almacenando datos masivos con HDFS
11. Procesando grandes cantidades de datos con MapReduce.
12. Herramientas integradas de Hadoop: Hive, Pig y Mahout

Actividades Formativas:

Actividad Formativa	Horas	Presencialidad
Clases Expositivas	30	0%
Ejercicios prácticos	30	0%
Seminarios	10	0%
Estudio autónomo	70	0%
Tutoría	10	0%

Metodologías docentes:

- Aprendizaje basado en problemas
- Aprendizaje basado en la experiencia
- Trabajo directo sobre plataformas tecnológicas digitales
- Estudio de casos

Sistema de Evaluación:

Sistemas de evaluación	Ponderación mínima	Ponderación máxima
Exposiciones orales	10.0	20.0
Portafolio	10.0	30.0
Trabajos individuales dirigidos	20.0	40.0
Pruebas de conocimiento	40.0	60.0