Universitat de Lleida

GUÍA DOCENTE

SISTEMAS INTENSIVOS DE PROCESO DE DATOS

Coordinación: MATEU PIÑOL, CARLOS

Año académico 2022-23

SISTEMAS INTENSIVOS DE PROCESO DE DATOS 2022-23

Información general de la asignatura

<table>
<tr><td>Denominación</td><td>SISTEMAS INTENSIVOS DE PROCESO DE DATOS</td></tr>
<tr><td>Código</td><td>103088</td></tr>
<tr><td>Semestre de impartición</td><td>2o Q(SEMESTRE) EVALUACIÓN CONTINUADA</td></tr>
<tr><td>Carácter</td><td>
<table>
<tr><th>Grado/Máster</th><th>Curso</th><th>Carácter</th><th>Modalidad</th></tr>
<tr><td>Máster Universitario en Ingeniería Informática</td><td>1</td><td>OPTATIVA</td><td>Presencial</td></tr>
</table>
</td></tr>
<tr><td>Número de créditos de la asignatura (ECTS)</td><td>6</td></tr>
<tr><td>Tipo de actividad, créditos y grupos</td><td>
<table>
<tr><th>Tipo de actividad</th><th>PRALAB</th><th>TEORIA</th></tr>
<tr><td>Número de créditos</td><td>3</td><td>3</td></tr>
<tr><td>Número de grupos</td><td>1</td><td>1</td></tr>
</table>
</td></tr>
<tr><td>Coordinación</td><td>MATEU PIÑOL, CARLOS</td></tr>
<tr><td>Departamento/s</td><td>INFORMATICA E INGENIERIA INDUSTRIAL</td></tr>
<tr><td>Información importante sobre tratamiento de datos</td><td>Consulte este enlace para obtener más información.</td></tr>
<tr><td>Idioma/es de impartición</td><td>Anglés</td></tr>
</table>

SISTEMAS INTENSIVOS DE PROCESO DE DATOS 2022-23

Profesor/a (es/as)	Dirección electrónica\nprofesor/a (es/as)	Créditos impartidos por el profesorado	Horario de tutoría/lugar
CORES PRADO, FERNANDO	fernando.cores@udl.cat	3	
LAMAS PRIETO, ALBA	alba.lamas@udl.cat	3	

Objetivos académicos de la asignatura

- **Decide y diseña una arquitectura distribuida adecuada para dar respuesta a un problema que involucra Big Data. Elige una tecnología adecuada para implantar esa arquitectura.**
- **Conoce las aplicaciones habituales en que aparecen problemas en el ámbito del Big Data y es capaz de desarrollar soluciones a esos problemas.**
- **Desarrolla sistemáticamente un proyecto de resolución de un problema en uno de los ámbitos típicos del Big Data.**
- **Comunica eficazmente los resultados de su proyecto a interlocutores técnicos y a clientes.**

Competencias

Competecias Generales

CG4 Capacidad para el modelado matemático, cálculo y simulación en centros tecnológicos y de ingeniería de empresa, particularmente en tareas de investigación, desarrollo e innovación en todos los ámbitos relacionados con la Ingeniería en Informática.

CG8 Capacidad para la aplicación de los conocimientos adquiridos y de resolver problemas en entornos nuevos o poco conocidos dentro de contextos más amplios y mulitidisciplinares, siendo capaces de integrar estos conocimientos.

Competecias Básicas

CB3 Ser capaz de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios.

CB4 Saber comunicar las conclusiones -y los conocimientos y razones últimas que las sustentan- a públicos especializados y no especializados de un modo claro y sin ambigüedades.

Competencias Específicas

CE1. Capacidad para la integración de tecnologías, aplicaciones, servicios y sistemas propios de la Ingeniería Informática, con carácter generalista, y en contextos más amplios y multidisciplinares.

CE4. Capacidad para modelar, diseñar, definir la arquitectura, implantar, gestionar, operar, administrar y mantener aplicaciones, redes, sistemas, servicios y contenidos informáticos.

CE10. Capacidad para comprender y poder aplicar conocimientos avanzados de computación de altas prestaciones y métodos numéricos o computacionales a problemas de ingeniería.

CE12. Capacidad para aplicar métodos matemáticos, estadísticos y de inteligencia artificial para modelar, diseñar

y desarrollar aplicaciones, servicios, sistemas inteligentes y sistemas basados ??en el conocimiento.

Contenidos fundamentales de la asignatura

1. Part I - Data Gathering & formating
    1. Introduction
    2. Open Data & Linked Data
    3. Internet Data Collection
        1. Data providing APIs
        2. Data Streams
    4. IoT as a data source
    5. Data Crowdsourcing
    6. Main data formats
        1. JSON
        2. XML
    7. Data correcting and cleanliness
2. Part II - Data storage and processing
    1. **Hadoop**
        1. ?Why Hadoop?
        2. Hadoop Concepts
        3. Hadoop Use Cases
        4. Components and Architecture
            1. HDFS
            2. Hadoop 2.0
        5. Planning a Installing an Hadoop Cluster
        6. Case study: Installation and Configuration Hadoop
    2. MapReduce Paradigm
        1. MapReduce model.
        2. Anatomy of a MapReduce Job
        3. Map Function
        4. Reduce Function
        5. Configuring and running a MapReduce job
    3. **Introduction to Apache Spark**
        1. What is Spark?
        2. The Spark Programming model
        3. Using Spark's Shells
        4. Working with Resilient Distributed Datasets (RDDs)
        5. Programming with Spark
        6. Setting Up Spark

SISTEMAS INTENSIVOS DE PROCESO DE DATOS 2022-23

Ejes metodológicos de la asignatura

Todos los cursos del bloque Big Data Analytics (Incluyendo éste), serán calificados con un proyecto único, común, que involucre todos los temas de los cursos (recopilación de datos, procesamiento, aprendizaje, estadísticas, visualización, etc.). Desde el principio (este curso) hasta los cursos finales.

Durante los cursos regulares, se introducirán diferentes temas, mostrando su relación con el proyecto común y cómo todos los temas encajan para crear una tarea o proyecto complejo del mundo real.

Los tres cursos que forman Big Data Analytics utilizarán la misma configuración de base tecnológica:

- Python como el lenguaje de programación base.

- Hadoop / Spark (con Java si es necesario)

- Aunque durante los cursos se introducirán otras suites tecnológicas: Scala, NodeJS, MongoDB, etc., según el tiempo lo permita.

Plan de desarrollo de la asignatura

Week	Description	Classroom Activity	Autonomous work Activity
1	Course introduction and preliminaries	Presentation Subject	Work Group Seminar
2	Data Gathering and Collection	Data Gathering and Collection	Bibliography and program review Preparing Project Idea
3	Data Gathering and Collection	Data Gathering and Collection	Preparing Project Idea
4	Data Gathering and Collection	Data Gathering and Collection	Big Data Project: Data Gathering
5	Data Cleansing and Conversion	Data Cleansing and Conversion	Big Data Project: Data Gathering
6	Data Cleansing and Conversion	Data Cleansing and Conversion	Big Data Project: Data cleaning
7	Data Cleansing and Conversion	Data Cleansing and Conversion	Big Data Project: Data cleaning
8	Hadoop Introduction	Hadoop Concepts & Use Cases	Study Hadoop Ecosystem
9	Hadoop Introduction	Hadoop Components and Architecture installation	HDFS Tutorial
10	MapReduce Paradigm	MapReduce model.	Big Data Project
11	MapReduce Paradigm	Anatomy of a MapReduce Job	Big Data Project MapReduce Tutorial
12	MapReduce Paradigm	Programing, configuring and running a MapReduce job	Big Data Project MapReduce Tutorial
13	Introduction to Spark	The Spark Programming model	Big Data Project Spark Tutorial
14	Introduction to Spark	Using Spark's Shells	Big Data Project Spark Tutorial
15	Introduction to Spark	Programming Spark and RDDs	Big Data Project Spark Tutorial
16	Final Project Delivery	**BigData Project Delivery**	

17	Project presentation	**BigData Project Presentation**
18		
19		

Sistema de evaluación

Acr.	Actividad	Peso	Puntos mínimos	Grupal?	Obligatoria	Recuperable
P1	Laboratorio parte 1	25%	NO	2-3	Si	NO
P2	Laboratorio parte 2	25%	NO	NO	Si	NO
PR	BigData Project 1	50%	NO	2	Si	NO

Las dos partes (I & II) se evaluaran:

- 50% de la evaluación será del proyecto global.

- 50% será de assignaciones específicas de la parte 1 y 2.

Bibliografía y recursos de información

Bibliografía básica:

[Whi15] Tom White, "Hadoop: The Definitive Guide", O'Reilly, 2015

[Hol15] Alex Holmes, "Hadoop in Practice", Manning, 2015.

[Kar15] Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, "Learning Spark: Lightning-Fast Big Data Analysis", O'Reilly, 2015

[Mar15] Nathan Marz, James Warren, "Big Data: Principles and best practices of scalable realtime data systems", Manning, 2015.

Bibliografía extendida:

[Ven14] Jason Venner, Sameer Wadkar, Madhu Siddalingaiah, "Pro Apache Hadoop", Apress, 2014.

[Bae14] Bart Baesens, "Analytics in a Big Data World: The Essential Guide to Data Science and its Applications"

[Ryz15] Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills, "Advanced Analytics with Spark: Patterns for Learning from Data at Scale", O'Reilly, 2015

[Gun15] Thilina Gunarathne, "Hadoop MapReduce Cookbook", 2015