Universitat de Lleida

GUIA DOCENT

SISTEMES INTENSIUS DE PROCESSAMENT DE DADES

Coordinació: PLANES CID, JORDI

Any acadèmic 2023-24

Informació general de l'assignatura

<table>
<tr><td>Denominació</td><td>SISTEMES INTENSIUS DE PROCESSAMENT DE DADES</td></tr>
<tr><td>Codi</td><td>103088</td></tr>
<tr><td>Semestre d'impartició</td><td>2N Q(SEMESTRE) AVALUACIÓ CONTINUADA</td></tr>
<tr><td>Caràcter</td><td>
<table>
<tr><th>Grau/Màster</th><th>Curs</th><th>Caràcter</th><th>Modalitat</th></tr>
<tr><td>Màster Universitari en Enginyeria Informàtica</td><td>1</td><td>OPTATIVA</td><td>Presencial</td></tr>
</table>
</td></tr>
<tr><td>Nombre de crèdits assignatura (ECTS)</td><td>6</td></tr>
<tr><td>Tipus d'activitat, crèdits i grups</td><td>
<table>
<tr><th>Tipus d'activitat</th><th>PRALAB</th><th>TEORIA</th></tr>
<tr><td>Nombre de crèdits</td><td>3</td><td>3</td></tr>
<tr><td>Nombre de grups</td><td>1</td><td>1</td></tr>
</table>
</td></tr>
<tr><td>Coordinació</td><td>PLANES CID, JORDI</td></tr>
<tr><td>Departament/s</td><td>ENGINYERIA INFORMÀTICA I DISSENY DIGITAL</td></tr>
<tr><td>Informació important sobre tractament de dades</td><td>Consulteu aquest enllaç per a més informació.</td></tr>
<tr><td>Idioma/es d'impartició</td><td>Anglés</td></tr>
</table>

Professor/a (s/es)	Adreça electrònica professor/a (s/es)	Crèdits impartits pel professorat	Horari de tutoria/lloc
LAMAS PRIETO, ALBA	alba.lamas@udl.cat	4,5	
PLANES CID, JORDI	jordi.planes@udl.cat	1,5	

Objectius acadèmics de l'assignatura

- Decideix i dissenya una arquitectura distribuïda adequada per donar resposta a un problema que involucra Big Data. Tria una tecnologia adequada per implantar aquesta arquitectura.
- Coneix les aplicacions habituals en què apareixen problemes en l'àmbit del Big Data i és capaç de desenvolupar solucions a aquests problemes.
- Desenvolupa sistemàticament un projecte de resolució d'un problema en un dels àmbits típics del Big Data.
- Comunica eficaçment els resultats del seu projecte a interlocutors tècnics i als clients.

Competències

Competècies Generals

CG4 Capacitat per al modelatge matemàtic, càlcul i simulació en centres tecnològics i d'enginyeria d'empresa, particularment en tasques d'investigació, desenvolupament i innovació en tots els àmbits relacionats amb l'Enginyeria en Informàtica.

CG8 Capacitat per a l'aplicació dels coneixements adquirits i de resoldre problemes en entorns nous o poc coneguts dins de contextos més amplis i mulitidisciplinares, sent capaços d'integrar aquests coneixements.

Competècies Bàsiques

CB3 Ser capaç d'integrar coneixements i enfrontar-se a la complexitat de formular judicis a partir d'una informació que, sent incompleta o limitada, inclogui reflexions sobre les responsabilitats socials i ètiques vinculades a l'aplicació dels seus coneixements i judicis.

CB4 Saber comunicar les conclusions -i els coneixements i raons últimes que les sustenten- a públics especialitzats i no especialitzats d'una manera clara i sense ambigüitats.

Competències Específiques

CE1. Capacitat per a la integració de tecnologies, aplicacions, serveis i sistemes propis de l'Enginyeria Informàtica, amb caràcter generalista, i en contextos més amplis i multidisciplinaris.

CE4. Capacitat per modelar, dissenyar, definir l'arquitectura, implantar, gestionar, operar, administrar i mantenir aplicacions, xarxes, sistemes, serveis i continguts informàtics.

CE10. Capacitat per comprendre i poder aplicar coneixements avançats de computació d'altes prestacions i mètodes numèrics o computacionals a problemes d'enginyeria.

CE12. Capacitat per aplicar mètodes matemàtics, estadístics i d'intel·ligència artificial per modelar, dissenyar i desenvolupar aplicacions, serveis, sistemes intel·ligents i sistemes basats en el coneixement.

Continguts fonamentals de l'assignatura

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. CSV
        2. JSON
        3. 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

Eixos metodològics de l'assignatura

Tots els cursos del bloc Big Data Analytics (incloent aquest), seran avaluats per un únic projecte comú que involucri tots els temes del bloc (recopilació de dades, processament, l'aprenentatge automàtic, estadística, visualització, etc.) Els estudiants treballaran en aquest projecte des del primer curs (aquest) fins els cursos finals.

Durant els cursos regulars, s'introduiran diferents temes, mostrant-ne la seva relació amb el projecte comú i estudiant com tots els temes encaixen entre si per crear una tasca o un projecte complex en el món real.

Els tres cursos de formació de Big Data Analytics utilitzaran la mateixa configuració de base tecnològica:

- Python com a llenguatge de programació de base.

- Hadoop / Spark (amb Java, si cal)

- Tot i que durant els cursos es poden introduir altres paquets tecnològics: Scala, nodejs, MongoDB, etc. si el temps ho permet.

Pla de desenvolupament de l'assignatura

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 d'avaluació

Acr.	Activitat	Pes	Punts mínims	Grupal?	Obligatoria	Recuperable
P1	Laboratori part 1	25%	NO	NO	Si	NO
P2	Laboratori part 2	25%	NO	NO	Si	NO
PR	BigData Project 1	50%	NO	2	Si	Sí

Les dos parts (I & II) s'evaluaran:

- 50% de la evaluació serà del projecte global.

- 50% serà d'assignacions específiques de part 1 i 2.

Bibliografia i recursos d'informació

Bibliografia Bàsica:

[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.

Bibliografia Ampliada:

[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