DEGREE CURRICULUM

MASSIVE DATA PROCESSING

Coordination: MATEU PIÑOL, CARLOS

Academic year 2020-21

MASSIVE DATA PROCESSING 2020-21

Subject's general information

<table>
<tr><td>Subject name</td><td>MASSIVE DATA PROCESSING</td></tr>
<tr><td>Code</td><td>103088</td></tr>
<tr><td>Semester</td><td>2nd Q(SEMESTER) CONTINUED EVALUATION</td></tr>
<tr><td>Typology</td><td>
<table>
<tr><th>Degree</th><th>Course</th><th>Character</th><th>Modality</th></tr>
<tr><td>Master's Degree in Informatics Engineering</td><td>1</td><td>OPTIONAL</td><td>Attendance-based</td></tr>
</table>
</td></tr>
<tr><td>Course number of credits (ECTS)</td><td>6</td></tr>
<tr><td>Type of activity, credits, and groups</td><td>
<table>
<tr><th>Activity type</th><td>PRALAB</td><td>TEORIA</td></tr>
<tr><th>Number of credits</th><td>3</td><td>3</td></tr>
<tr><th>Number of groups</th><td>1</td><td>1</td></tr>
</table>
</td></tr>
<tr><td>Coordination</td><td>MATEU PIÑOL, CARLOS</td></tr>
<tr><td>Department</td><td>COMPUTER SCIENCE AND INDUSTRIAL ENGINEERING</td></tr>
<tr><td>Important information on data processing</td><td>Consult this link for more information.</td></tr>
<tr><td>Language</td><td>Anglés</td></tr>
</table>

MASSIVE DATA PROCESSING 2020-21

Teaching staff	E-mail addresses	Credits taught by teacher	Office and hour of attention
CORES PRADO, FERNANDO	fernando.cores@udl.cat	3	
MATEU PIÑOL, CARLOS	carles.mateu@udl.cat	3	

Learning objectives

- Design a suitable distributed architecture to address a problem involving Big Data. Select a suitable technology to implement that architecture.
- He is aware of the usual problems appearing on the field and the suitable applications of those, and is able to develop solutions to those problems.
- Is able to systematically develop a project solving a problem in one of the typical areas of Big Data.
- Effectively communicates the results of the project to technical partners and clients.

Competences

General Competences.

CG4 Capacity for mathematical modeling, calculation and simulation in technologic and engineering business centers, particularly in research tasks, development and innovation in all areas related to Computer Engineering.

CG8 Capacity to apply the acquired knowledge and solve problems in new or unfamiliar environments within broader contexts and mulitidisciplinares, being able to integrate this knowledge.

Basic Competences.

CB3 Being able to integrate knowledge and handle the complexity to formulate judgments based on information that being incomplete or limited, include reflecting on social and ethical responsibilities linked to the application of their knowledge and judgments.

CB4 Knowing how to communicate their conclusions -and the knowledge and rationale underpinning these, to specialist and non-specialist audiences clearly and unambiguously

Specific competences.

CE1. Capacity for the integration of technologies, applications and computer engineering systems, in general and in wider and multidisciplinary contexts.

CE4. Capacity to model, design, define the architecture, implant, manage, operate, administer and keep applications, networks, systems, services and computer contents.

CE10. Capacity to understand and apply advanced knowledge in high-performance computing and numerical or computational methods to problems of engineering.

CE12. Capacity to apply mathematical, statistical and artificial intelligence methods, design and develop applications, services, intelligent systems and systems based on knowledge.

MASSIVE DATA PROCESSING 2020-21

Subject contents

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

Methodology

MASSIVE DATA PROCESSING 2020-21

All the courses of the Big Data Analytics block (Including this one), will be graded by an unique, common, project involving all the courses topics (data gathering, processing, learning, statistics, visualization, etc.) Students will work on that project from the beginning (this course) to the final courses.

During regular courses, different topics will be introduced, showing their relation to the common project and how all topics fit together to create a real-world complex task or project.

The three courses forming Big Data Analytics will use the same technological base setup:

- Python as the base programming language.

- Hadoop/Spark (with Java if required)

- Although during the courses other technological suites will be introduced: Scala, NodeJS, MongoDB, etc. as time permits.

Development plan

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

Evaluation

Acr.	Evaluation Activity	Weight	Minimum Score	Group Work	Compulsory	Recoverable
P1	Lab work part 1	25%	NO	2-3	Yes	NO
P2	Lab work part 2	25%	NO	NO	Yes	NO
PR	BigData Project Deliverable 1	50%	Yes	2	Yes	NO

Both parts (I & II) will be evaluated as follows:

- 50% of the evaluation will be for the global Big Data Analytics project.

- 50% will be on specialized assignments for both Part I and Part II.

Bibliography

Basic Bibliography:

[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.

Extended Bibliography:

[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