Teaching Guide				
Identifying Data				**2019/20**
Subject (*)	Data Analytics with HPC		**Code**	614973108
Study programme	Mestrado Universitario en Computación de Altas Prestacións / High Performance Computing (Mod. Virtual)			
Descriptors				
Cycle	**Period**	**Year**	**Type**	**Credits**
Official Master's Degree	2nd four-month period	First	Optional	6
Language	English			
Teaching method	Face-to-face			
Prerequisites				
Department	Departamento profesorado másterEnxeñaría de Computadores			
Coordinador	López Taboada, Guillermo	**E-mail**	guillermo.lopez.taboada@udc.es	
Lecturers	López Taboada, Guillermo Rodríguez Álvarez, Gabriel	**E-mail**	guillermo.lopez.taboada@udc.es gabriel.rodriguez@udc.es	
Web	aula.cesga.es			
General description	The increasing amount of information available through the Internet calls for the efficient processing of large amounts of data. This has led to the development of new storage and processing techniques to deal with huge amounts of data, namely Big Data techniques, that naturally adapt to distributed systems. The main goal of this subject is to learn suitable processing techniques for large amounts of information in the Big Data world, particularly using the Hadoop ecosystem, and compare these techniques with the traditional ones employed in HPC environments. This will allow the student to select the optimal tools to solve a particular problem.			

Study programme competences	
Code	**Study programme competences**
A1	CE1 - Define, evaluate and select the most appropriate architecture and software to solve a problem
A2	CE2 - Analyze and improve the performance of a given architecture or software
B1	CB6 - Possess and understand the knowledge that give a baseline or opportunity to be original in the development and/or application of ideas, often in a research environment
B2	CB7 - The students have to know how to apply the acquired knowledge and their capacity to solve problems in new or hardly explored environment inside wider contexts (or multidiscipinary) related to its area of development
B6	CG1 - Be able to search and select useful information to solve complex problems, using the bibliographic sources of the field
B8	CG3 - Be able to maintain and extend properly funded theoretical hypothesis to allow the introduction and exploitation of novel and advanced technologies in the field
B10	CG5 - Be able to work in teams, specially multidisciplinary, and do a proper time and people management and decision taking
C1	CT1 - Use the basic technologies of the information and computing technology field required for the professional development and the long-life learning
C4	CT4 - Value the importance of research, innovation and the technological development in the socioeconomical and cultural advance of the society

Learning outcomes			
Learning outcomes	**Study programme competences**		
The student will be capable of installing, configuring, and managing the basic software for massive data processing.	AJ1 AJ2	BJ2 BJ6 BJ8 BJ10	CJ1
The student will be capable of coding massive data processing applications using domain-specific languages.	AJ2	BJ1 BJ2 BJ10	CJ1

The student will learn about Data Engineering tools (for Intake/Storage/Processing/Visualization).	AJ1 AJ2	BJ1 BJ2	CJ1 CJ4
The student will learn the skills to search, select and manage Big data-related resources (bibliography, software, etc.).	AJ1 AJ2	BJ1 BJ6	CJ1 CJ4

Contents	
Topic	**Sub-topic**
1. Introduction to Data Engineering	1.1 HPC vs Big Data: similarities and differences in data management. 1.2 Hardware and Software Technologies for High Performance Data Engineering 1.3 Data Engineering in HPC infrastructures vs. Cloud environments
2. Data Engineering phases	2.1 Modeling (Formats, Compression, Designing Schemas) 2.2 Intake (Periodicity, Transformations, Tools) 2.3 Storage (HDFS and NoSQL DBs, HBase, MongoDB, Cassandra) 2.4 Processing (Batch, Real-Time) 2.5 Orchestration 2.6 Analysis (SQL, Machine Learning, Graphs, UI) 2.7 Governance 2.8 Integration with BI (Visualization)
3. Introduccion to Data Analytics	3.1 Exploratory Data Analytics 3.2 Introduction to Machine Learning
4 Use cases	4.1 Applications to Internet of Things (Smart environments and Industry 4.0) 4.2 Applications to sciences and engineering

Planning				
Methodologies / tests	**Competencies**	**Ordinary class hours**	**Student?s personal work hours**	**Total hours**
Workbook	A1 A2 B1 B6 C4	0	18	18
Laboratory practice	B1 B8 B10	0	80	80
Supervised projects	A1 A2 B1 B2 B8	0	45	45
Directed discussion	B6 C1 C4	4	2	6
Personalized attention		1	0	1
(*)The information in the planning table is for guidance only and does not take into account the heterogeneity of the students.				

Methodologies	
Methodologies	**Description**
Workbook	Planned instruction through various teaching materials.
Laboratory practice	Problem solving and practical cases.
Supervised projects	Semi-autonomous work on larger practical cases, under the professors' guidance.
Directed discussion	Guidance to solve individual / group assignments, problem solving and continuous evaluation activities.

Personalized attention	
Methodologies	**Description**
Laboratory practice Supervised projects Directed discussion	During laboratory practice, supervised projects, and directed discussions, students will be able to ask questions, doubts, etc. The teacher, after listening to the students feedback, will go over difficult concepts, solve new problems, or use any appropriate methodology to answer the questions.

Assessment			
Methodologies	**Competencies**	**Description**	**Qualification**

Laboratory practice	B1 B8 B10	Grading the assignments submitted by students.	40
Supervised projects	A1 A2 B1 B2 B8	Grading the supervised projects submitted by students.	50
Directed discussion	B6 C1 C4	Continued, active, objectively measurable participation by the student.	10

Assessment comments

First evaluation (May):

Practical exercises: 40% Guided projects: 50% Objective participation: 10% Second evaluation (June/July):

Practical exercises: same grade as in the first evaluation, as there are no new activities planned for this evaluation. 40% of the final grade. Guided projects: projects not evaluated in may or deemed incomplete will be presented in july after performing the changes suggested by the professor. 50% of the final grade. Objective participation: same grade as in the first evaluation, as there are no new activities planned for this evaluation. 10% of the final grade. Not graded: Students that do not present any practical exercise or guided project will not be graded.

Sources of information

Basic	- Tom White (2015). Hadoop: The Definitive Guide. O'Reilly (4ª ed.) - Wes McKinney (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly (2ª ed.)
Complementary	- Alex Holmes (2014). Hadoop in practice. Manning (2ª ed.)

Recommendations

Subjects that it is recommended to have taken before

Subjects that are recommended to be taken simultaneously

Subjects that continue the syllabus

Other comments

RecommendationsDue to the large practical component of the subject, it is advisable to be up-to-date with practices and guided projects during the semester.&nbsp;Observations&nbsp;The course makes intensive use of online communication tools: Video calls, chats, etc. In-person classes will be recorded for later perusing. An online learning management will be using for distributing notes, creating forums, etc. &nbsp;The software tools used in this course are generally open-source or have free license for students.

(*)The teaching guide is the document in which the URV publishes the information about all its courses. It is a public document and cannot be modified. Only in exceptional cases can it be revised by the competent agent or duly revised so that it is in line with current legislation.