ANX-PR/CL/001-01

LEARNING GUIDE

SUBJECT

103000893 - Big Data

DEGREE PROGRAMME

10BA - Master Universitario En Ciencia De Datos

ACADEMIC YEAR & SEMESTER

2022/23 - Semester 1

Index

Learning guide

1. Description

1.1. Subject details

Name of the subject	103000893 - Big Data
No of credits	3 ECTS
Type	Compulsory
Academic year ot the programme	First year
Semester of tuition	Semester 1
Tuition period	September-January
Tuition languages	English
Degree programme	10BA - Master Universitario en Ciencia de Datos
Centre	10 - Escuela Tecnica Superior De Ingenieros Informaticos
Academic year	2022-23

2. Faculty

2.1. Faculty members with subject teaching role

Name and surname	Office/Room	Email	Tutoring hours *
Antonio Latorre De La Fuente (Subject coordinator)	4202	a.latorre@upm.es	Sin horario.
Jesus Montes Sanchez	4204	jesus.montes@upm.es	Sin horario.

* The tutoring schedule is indicative and subject to possible changes. Please check tutoring times with the faculty member in charge.

3. Skills and learning outcomes *

3.1. Skills to be learned

CB07 - Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio

CB10 - Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

CECD02 - Conocer los principales sistemas de almacenamiento de la información estructurada y no estructurada de fuentes heterogéneas.

CECD03 - Manejar las herramientas informáticas para Big Data

CG07 - Aplicación de los últimos o más novedosos métodos para resolver problemas que, posiblemente, involucren a otras disciplinas

CG10 - Apreciación de los límites del conocimiento actual y de la aplicación práctica de la última tecnología

CG11 - Conocimiento y comprensión de la informática para crear modelos, así como sistemas y procesos de información complejos

3.2. Learning outcomes

RA20 - Ser capaz de procesar datos masivos

RA21 - Conocer cómo se aplican las técnicas de computación científica en algún campo específico de ciencia o ingeniería

RA19 - Conocer técnicas y procesos de análisis de datos, y de programación, diseño y depuración de algoritmos, para computación de altas prestaciones

* The Learning Guides should reflect the Skills and Learning Outcomes in the same way as indicated in the Degree Verification Memory. For this reason, they have not been translated into English and appear in Spanish.

4. Brief description of the subject and syllabus

4.1. Brief description of the subject

This course will allow the student to gain the fundamentals for the analysis of large volumes of data. With an eminently practical approach, the technologies and fundamentals necessary to successfully accomplish the whole data analysis process will be presented in the context of Big Data, from the raw data to the models derived from them.

4.2. Syllabus

1. Introduction to Big Data
    1.1. Architectures and applications
    1.2. Data types
    1.3. Visual analytics
2. Big Data Ecosystem
3. Big Data Technologies
    3.1. Technological Challenges
    3.2. Basic solution: gfs + MapReduce
    3.3. Hadoop (hdfs + yarn)
    3.4. Pig
    3.5. Hive
    3.6. Beyond MapReduce
        3.6.1. Tez
        3.6.2. Spark
        3.6.3. Flink
4. Spark
    4.1. Spark Basics
    4.2. Brief Introduction to Scala
    4.3. Spark Applications

4.4. Spark SQL

5. Machine Learning with Spark

5.1. Brief review of Machine Learning basics

5.2. Spark MLlib

5. Schedule

5.1. Subject schedule*

Week	Classroom activities	Laboratory activities	Distant / On-line	Assessment activities
1	**Lesson 1** Duration: 02:00 **Lesson 2** Duration: 01:00			
2	**Lesson 2** Duration: 01:00 **Lesson 2** Duration: 02:00			
3	**Lesson 3** Duration: 01:00	**Practical Work** Duration: 02:00		
4	**Lesson 3** Duration: 01:00	**Practical Work** Duration: 02:00		
5	**Lesson 4** Duration: 01:00 **Lesson 4** Duration: 01:00	**Practical Work** Duration: 01:00		
6	**Lesson 4** Duration: 01:00 **Lesson 4** Duration: 01:00	**Practical Work** Duration: 01:00		
7	**Lesson 4** Duration: 01:00	**Practical Work** Duration: 02:00		
8	**Lesson 4** Duration: 01:00	**Practical Work** Duration: 02:00		
9	**Lesson 5** Duration: 01:00 **Lesson 5** Duration: 01:00	**Practical Work** Duration: 01:00		

10				
11				
12				
13				
14				
15				
16				**Assignment Deadline** Continuous assessment and final examination Not Presential Duration: 00:00
17				**Final Exam** Continuous assessment and final examination Presential Duration: 02:00

Depending on the programme study plan, total values will be calculated according to the ECTS credit unit as 26/27 hours of student face-to-face contact and independent study time.

* The schedule is based on an a priori planning of the subject; it might be modified during the academic year, especially considering the COVID19 evolution.

6. Activities and assessment criteria

6.1. Assessment activities

6.1.1. Assessment

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
16	Assignment Deadline		No Presential	00:00	80%	4 / 10	CECD02 CG07 CECD03 CB10 CG11 CB07 CG10
17	Final Exam		Face-to-face	02:00	20%	4 / 10	CECD02 CG07 CECD03 CB10 CG11 CB07 CG10

6.1.2. Global examination

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
16	Assignment Deadline		No Presential	00:00	80%	4 / 10	CECD02 CG07 CECD03 CB10 CG11 CB07 CG10
17	Final Exam		Face-to-face	02:00	20%	4 / 10	CECD02 CG07 CECD03 CB10 CG11 CB07 CG10

6.1.3. Referred (re-sit) examination

No se ha definido la evaluación extraordinaria.

6.2. Assessment criteria

Regular call

This section covers the evaluation criteria for this course. All the students enrolled in this course will be subject, by default, to the progressive evaluation scheme. For this reason, this learning guide will be focused on this evaluation approach and it details all the evaluation activities in the timeline of the course.

The evaluation of the course will take into account both the theoretical and the practical knowledge acquired in the lectures and in the practical work carried out during the course, respectively.

This course will be evaluated in two ways:

- **Final exam.** At the end of the course, there will be a final exam covering all the contents presented during the course.
- **Practical work.** This assignment will be presented during the course, at class, in the date detailed in the timeline of the course. There will be some classes devoted to this assignment, in which the students will count with the support of the instructor, which should be, in general, complemented with autonomous work by the student. The deadline for the assignment will be fixed at the end of the term, as shown in the timeline of the course. No late assignments will be accepted for evaluation. The delivery of the assignment is considered a mandatory activity in order to pass the course.

The **final grade** for this course will be computed as follows: 20% for the final exam and 80% for the assignment. To pass the course, **a minimum score of 4** is required for each of these parts and a **grand mean** of 5 is needed combining these two items of evaluation.

Extraordinary call

If the students do not succeed in this course, they will have to repeat the evaluation actvities not passed in the

ordinary evaluation. There will be a new call for the final exam as well as a new deadline for the assignment with the same requirements as in the regular call.

7. Teaching resources

7.1. Teaching resources for the subject

Name	Type	Notes
Book 1	Bibliography	Jiawei Han, Micheline Kamber, Data Mining : Concepts and Techniques, 2nd edition, Morgan Kaufmann, ISBN 1558609016, 2006.
Book 2	Bibliography	Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson Addison Wesley, ISBN: 0321321367, 2005
Book 3	Bibliography	Ian Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufmann, ISBN: 0120884070, 2005.
Book 4	Bibliography	Ian Witten, Eibe Frank, Mark Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3nd Edition, Morgan Kaufmann, ISBN: 978-0-12-374856-0, 2011.
Book 5	Bibliography	Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia. Learning Spark: Lightning-Fast Big Data Analysis. O'Reilly Media. 2015.
Book 6	Bibliography	Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills. Advanced Analytics with Spark: Patterns for Learning from Data at Scale. O'Reilly Media. 2015.

Spark documentation	Web resource	http://spark.apache.org/docs/latest/
Web site of the course	Web resource	UPM Moodle
Hive documentation	Web resource	https://cwiki.apache.org/confluence/display/Hive/Home

8. Other information

8.1. Other information about the subject

This course is jointly offered with other Master Programmes and lectures are delivered in English.