



POLITÉCNICA

INTERNATIONAL  
CAMPUS OF  
EXCELLENCE

COORDINATION PROCESS OF  
LEARNING ACTIVITIES  
PR/CL/001



E.T.S. de Ingenieros  
Informáticos

# ANX-PR/CL/001-01

## LEARNING GUIDE

### SUBJECT

**103000893 - Big Data**

### DEGREE PROGRAMME

10BA - Master Universitario En Ciencia De Datos

### ACADEMIC YEAR & SEMESTER

2021/22 - Semester 1

## Index

---

### Learning guide

1. Description.....	1
2. Faculty.....	1
3. Skills and learning outcomes .....	2
4. Brief description of the subject and syllabus.....	3
5. Schedule.....	5
6. Activities and assessment criteria.....	7
7. Teaching resources.....	10
8. Other information.....	11

## 1. Description

---

### 1.1. Subject details

<b>Name of the subject</b>	103000893 - Big Data
<b>No of credits</b>	3 ECTS
<b>Type</b>	Compulsory
<b>Academic year of the programme</b>	First year
<b>Semester of tuition</b>	Semester 1
<b>Tuition period</b>	September-January
<b>Tuition languages</b>	English
<b>Degree programme</b>	10BA - Master Universitario en Ciencia de Datos
<b>Centre</b>	10 - Escuela Tecnica Superior De Ingenieros Informaticos
<b>Academic year</b>	2021-22

## 2. Faculty

---

### 2.1. Faculty members with subject teaching role

<b>Name and surname</b>	<b>Office/Room</b>	<b>Email</b>	<b>Tutoring hours *</b>
Antonio Latorre De La Fuente (Subject coordinator)	4202	a.latorre@upm.es	Sin horario.
Pablo Toharia Rabasco	4102	pablo.toharia@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

CB06 - Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación

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

CB08 - Que los estudiantes sean capaces 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

CB09 - Que los estudiantes sepan comunicar sus conclusiones y los conocimientos y razones últimas que las sustentan a públicos especializados y no especializados de un modo claro y sin ambigüedades

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

CG06 - Especificación y realización de tareas informáticas complejas, poco definidas o no familiares

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

CG08 - Capacidad de pensamiento de forma creativa para desarrollar aproximaciones y métodos nuevos y originales

CG09 - Integración del conocimiento de distintos campos de estudio

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

CGI01 - Adquirir conocimientos científicos avanzados del campo de la informática que le permitan generar nuevas ideas dentro de una línea de investigación

CGI02 - Comprender el procedimiento, valor y límites del método científico en el campo de la Informática, siendo capaz de identificar, localizar y obtener datos requeridos en un trabajo de investigación, de diseñar y guiar investigaciones analíticas, de modelado y experimentales, así como de evaluar datos de una manera crítica y extraer conclusiones

CGI03 - Capacidad para valorar la importancia de las fuentes documentales, manejarlas y buscar la información para el desarrollo de cualquier trabajo de investigación

CGI04 - Capacidad de leer y comprender publicaciones dentro de su ámbito de estudio/investigación, así como su catalogación y valor científico

### 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	Face-to-face classroom activities	Face-to-face laboratory activities	Distant / On-line	Assessment activities
1	<b>Lesson 1</b> Duration: 02:00  <b>Lesson 2</b> Duration: 01:00		<b>Lesson 1</b> Duration: 02:00  <b>Lesson 2</b> Duration: 01:00	
2	<b>Lesson 2</b> Duration: 01:00  <b>Lesson 2</b> Duration: 02:00		<b>Lesson 2</b> Duration: 01:00  <b>Lesson 2</b> Duration: 02:00	
3	<b>Lesson 3</b> Duration: 01:00	<b>Practical Work</b> Duration: 02:00	<b>Lesson 3</b> Duration: 01:00  <b>Practical Work</b> Duration: 02:00	
4	<b>Lesson 3</b> Duration: 01:00	<b>Practical Work</b> Duration: 02:00	<b>Lesson 3</b> Duration: 01:00  <b>Practical Work</b> Duration: 02:00	
5	<b>Lesson 4</b> Duration: 01:00  <b>Lesson 4</b> Duration: 01:00	<b>Practical Work</b> Duration: 01:00	<b>Lesson 4</b> Duration: 01:00  <b>Lesson 4</b> Duration: 01:00  <b>Practical Work</b> Duration: 01:00	
6	<b>Lesson 4</b> Duration: 01:00  <b>Lesson 4</b> Duration: 01:00	<b>Practical Work</b> Duration: 01:00	<b>Lesson 4</b> Duration: 01:00  <b>Lesson 4</b> Duration: 01:00  <b>Practical Work</b> Duration: 01:00	

7	Lesson 4 Duration: 01:00	Practical Work Duration: 02:00	Lesson 4 Duration: 01:00  Practical Work Duration: 02:00	
8	Lesson 4 Duration: 01:00	Practical Work Duration: 02:00	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	Lesson 5 Duration: 01:00  Lesson 5 Duration: 01:00  Practical Work Duration: 01:00	
10				Assignment Deadline  Continuous assessment and final examination Not Presential Duration: 00:00
11				
12				
13				
14				
15				
16				
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. Continuous assessment

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
10	Assignment Deadline		No Presential	00:00	80%	4 / 10	CB06 CG06 CG07 CG08 CG09 CG11 CECD02 CG10 CB08 CB07 CB09 CB10 CGI01 CGI04 CGI02 CGI03 CECD03
17	Final Exam		Face-to-face	02:00	20%	4 / 10	CB06 CG06 CG07 CG08 CG09 CG11 CECD02 CG10 CB08 CB07 CB09 CB10 CGI01 CGI04 CGI02 CGI03 CECD03

#### 6.1.2. Final examination

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
10	Assignment Deadline		No Presential	00:00	80%	4 / 10	CB06 CG06 CG07 CG08 CG09 CG11 CECD02 CG10 CB08 CB07 CB09 CB10 CGI01 CGI04 CGI02 CGI03 CECD03
17	Final Exam		Face-to-face	02:00	20%	4 / 10	CB06 CG06 CG07 CG08 CG09 CG11 CECD02 CG10 CB08 CB07 CB09 CB10 CGI01 CGI04 CGI02 CGI03 CECD03

### 6.1.3. Referred (re-sit) examination

No se ha definido la evaluación extraordinaria.

## 6.2. Assessment criteria

### Schedule

This schedule is preliminary and subject to changes that will be communicated on the website of the course.

### Continuous evaluation

This section covers the evaluation criteria for this course. All the students enrolled in this course will be subject, by default, to the continuous evaluation scheme. For this reason, this learning guide will be focused on this approach and details all the evaluation activities in the timeline of the course. Those students interested in the final examination evaluation scheme are referred to the next section of this document.

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 schedule of the course. There will be some classes devoted to this assignment, where the students will count with the support of the instructor, that should be, in general, complemented with autonomous work by the student. No late assignments will be accepted for evaluation.

The **final grade** for this 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.

### Final exam evaluation

In general, the regulations for this evaluation scheme will be the same as for the continuous evaluation option. In particular:

- The students will to conduct the same practical work without the in-class support of the instructors.
- The deadlines for the assignments will be the same as for the continuous evaluation scheme.

## Extraordinary evaluation in July

If the student does not succeed in this course, she will have to repeat those parts 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 of the course.

## 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, 3rd 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	<a href="http://spark.apache.org/docs/latest/">http://spark.apache.org/docs/latest/</a>
Web site of the course	Web resource	UPM Moodle
Hive documentation	Web resource	<a href="https://cwiki.apache.org/confluence/display/Hive/Home">https://cwiki.apache.org/confluence/display/Hive/Home</a>

## 8. Other information

---

### 8.1. Other information about the subject

This course is jointly offered with the EIT-Digital Master in Data Science and lectures are delivered in English.