ANX-PR/CL/001-01

LEARNING GUIDE

SUBJECT

103000640 - Big Data

DEGREE PROGRAMME

10AN - Master Universitario En Ingenieria Informatica

ACADEMIC YEAR & SEMESTER

2019/20 - Semester 1

Index

Learning guide

1. Description

1.1. Subject details

Name of the subject	103000640 - Big Data
No of credits	6 ECTS
Type	Optional
Academic year ot the programme	Second year
Semester of tuition	Semester 3
Tuition period	September-January
Tuition languages	English
Degree programme	10AN - Master Universitario En Ingenieria Informatica
Centre	10 - Escuela Tecnica Superior de Ingenieros Informaticos
Academic year	2019-20

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.
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. Prior knowledge recommended to take the subject

3.1. Recommended (passed) subjects

- Sistemas Cognitivos

- Sistemas Interactivos

- Computacion Para Ciencias E Ingenieria

3.2. Other recommended learning outcomes

The subject - other recommended learning outcomes, are not defined.

4. Skills and learning outcomes *

4.1. Skills to be learned

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.

CB7 - 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

CE1 - Capacidad para la integración de tecnologías, aplicaciones, servicios y sistemas propios de la Ingeniería Informática, con carácter generalista, y en contextos más amplios y multidisciplinares.

CE10 - Capacidad para comprender y poder aplicar conocimientos avanzados de computación de altas prestaciones y métodos numéricos o computacionales a problemas de ingeniería.

CE16 - Habilidad para hacer conexiones entre los deseos y necesidades del consumidor o cliente y lo que la tecnología puede ofrecer

CE4 - Capacidad para modelar, diseñar, definir la arquitectura, implantar, gestionar, operar, administrar y mantener aplicaciones, redes, sistemas, servicios y contenidos informáticos.

CE9 - Capacidad para diseñar y evaluar sistemas operativos y servidores, y aplicaciones y sistemas basados en computación distribuida.

CG12 - Capacidad de trabajar de forma independiente en su campo profesional

CG9 - Apreciación de los límites del conocimiento actual y de la aplicación práctica de la tecnología más reciente

4.2. Learning outcomes

RA41 - Conocer cómo se aplican las técnicas de computación científica en algún campo específico de ciencia o ingeniería

RA40 - Conocer técnicas de visualización y proceso de análisis de datos

RA108 - Conocer técnicas de visualización y procesos de análisis de datos, y de programación, diseño y depuración de algoritmos, para computación de altas prestaciones.

RA39 - Ser capaz de procesar datos masivos

* 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.

5. Brief description of the subject and syllabus

5.1. Brief description of the subject

This course will allow the student to gain the fundamentals for the analytical visualization 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 its visualization, through the models derived from them.

5.2. 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
6. Information Visualization
    6.1. Information Visualization Fundamentals
    6.2. Data Abstractions
    6.3. Tasks Abstractions

6.4. Interaction Techniques and Visual Encoding

6.5. Design Methods

6.6. Visualization Examples Analysis

6.7. Lessons Learnt

6. Schedule

6.1. Subject schedule*

Week	Face-to-face classroom activities	Face-to-face laboratory activities	Other face-to-face activities	Assessment activities
1	**Lesson 1** Duration: 02:00 Lecture **Lesson 2** Duration: 01:00 Lecture			
2	**Lesson 2** Duration: 01:00 Lecture **Lesson 2** Duration: 02:00 Problem-solving class			
3	**Lesson 3** Duration: 01:00 Lecture	**Practical Work 1** Duration: 02:00 Laboratory assignments		
4	**Lesson 3** Duration: 01:00 Problem-solving class	**Practical Work 1** Duration: 02:00 Laboratory assignments		
5	**Lesson 4** Duration: 01:00 Lecture **Lesson 4** Duration: 01:00 Problem-solving class	**Practical Work 1** Duration: 01:00 Laboratory assignments		
6	**Lesson 4** Duration: 01:00 Lecture **Lesson 4** Duration: 01:00 Problem-solving class	**Practical Work 1** Duration: 01:00 Laboratory assignments		
7	**Lesson 4** Duration: 01:00 Lecture	**Practical Work 1** Duration: 02:00 Laboratory assignments		
8	**Lesson 4** Duration: 01:00 Lecture	**Practical Work 1** Duration: 02:00 Laboratory assignments		
9	**Lesson 5** Duration: 01:00 Lecture **Lesson 5** Duration: 01:00 Problem-solving class	**Practical Work 1** Duration: 01:00 Laboratory assignments		**First Assignment Deadline** Group work Continuous assessment and final examination Duration: 00:00

10	**Lesson 6** Duration: 01:00 Lecture **Lesson 6** Duration: 01:00 Problem-solving class	**Practical Work 2** Duration: 01:00 Laboratory assignments		
11	**Lesson 6** Duration: 02:00 Lecture	**Practical Work 2** Duration: 01:00 Laboratory assignments		
12	**Lesson 6** Duration: 01:00 Lecture **Lesson 6** Duration: 01:00 Problem-solving class	**Practical Work 2** Duration: 01:00 Laboratory assignments		
13	**Lesson 6** Duration: 02:00 Lecture	**Practical Work 2** Duration: 01:00 Laboratory assignments		
14	**Lesson 6** Duration: 01:00 Lecture	**Practical Work 2** Duration: 02:00 Laboratory assignments		
15	**Lesson 6** Duration: 01:00 Lecture	**Practical Work 2** Duration: 02:00 Laboratory assignments		
16		**Practical Work 2** Duration: 03:00 Laboratory assignments		**Second Assignment Deadline** Group work Continuous assessment and final examination Duration: 00:00
17				**Final Exam** Written test Continuous assessment and final examination Duration: 01:00

The independent study hours are training activities during which students should spend time on individual study or individual assignments.

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 subject schedule is based on a previous theorical planning of the subject plan and might go to through experience some unexpected changes along throughout the academic year.

7. Activities and assessment criteria

7.1. Assessment activities

7.1.1. Continuous assessment

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
9	First Assignment Deadline	Group work	No Presential	00:00	40%	4 / 10	CB7 CB10 CE4 CG9 CG12 CE16 CE1 CE9 CE10
16	Second Assignment Deadline	Group work	No Presential	00:00	40%	4 / 10	CB7 CB10 CE4 CG9 CG12 CE16 CE1 CE9 CE10
17	Final Exam	Written test	Face-to-face	01:00	20%	4 / 10	CB7 CB10 CE4 CG9 CG12 CE16 CE1 CE9 CE10

7.1.2. Final examination

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
9	First Assignment Deadline	Group work	No Presential	00:00	40%	4 / 10	CB7 CB10 CE4 CG9 CG12 CE16 CE1 CE9

							CE10
16	Second Assignment Deadline	Group work	No Presential	00:00	40%	4 / 10	CB7 CB10 CE4 CG9 CG12 CE16 CE1 CE9 CE10
17	Final Exam	Written test	Face-to-face	01:00	20%	4 / 10	CB7 CB10 CE4 CG9 CG12 CE16 CE1 CE9 CE10

7.1.3. Referred (re-sit) examination

No se ha definido la evaluación extraordinaria.

7.2. Assessment criteria

Continuous evaluation

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. Al 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**. These assignments will be presented during the course, at class, in the dates detailed in the timeline of the course. There will be some classes devoted to these assignments, where the students will count with the support of the instructor, that should be, in general, complemented with autonomous work by the student. The deadlines for the assignment are spread through the term, as shown in the timeline of the course. No late assignments will be accepted for evaluation.

The **final grade** for this course will be computed as follows: 20% for the final exam, 40% for the first assignment, and, finally, 40% for the second 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 three items of evaluation.

Final exam evaluation

This evaluation scheme will be only offered if the current regulations of the UPM requires it and the procedure to opt for this type of evaluation will be subject to the instructions given by the school. Please, refer to **http://www.fi.upm.es/?pagina=1147** for additional information.

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

- The students will have to conduct the same practical works 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 common for all the assignments of the course.

8. Teaching resources

8.1. 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	Keim, D., Kohlhammer, J., Ellis, G., Mansmann, F. Mastering the information age. Solving problems with visual analytics 2010 Eurographics Association.
Book 6	Bibliography	Tamara Munzner. Visualization Analysis and Design. A K Peters Visualization Series. CRC Press. Nov. 2014.
Book 7	Bibliography	Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia. Learning Spark: Lightning-Fast Big Data Analysis. O'Reilly Media. 2015.
Book 8	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/
Assigned class	Equipment	
Web site of the course	Web resource	UPM Moodle
Hive documentation	Web resource	https://cwiki.apache.org/confluence/display/Hive/Home

9. Other information

9.1. Other information about the subject

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