COURSE DATA

Data Subject	
Code	44658
Name	Big Data
Cycle	Master's degree
ECTS Credits	6.0
Academic year	2020 - 2021

Study (s)

Degree	Center	Acad. year	Period
2221 - M.U. en Ciencia de Datos	School of Engineering	1	Second term

Subject-matter

Degree	Subject-matter	Character
2221 - M.U. en Ciencia de Datos	10 - Big data	Obligatory

Coordination

Name	Department
LAPARRA PEREZ-MUELAS, VALERO	242 - Electronic Engineering
LOZANO IBAÑEZ, MIGUEL	240 - Computer Science
SEBASTIAN AGUILAR, RAFAEL	240 - Computer Science

SUMMARY

This subject is organised in three main blocks. First block introduces the most common drawbacks that arise when processing huge data volumes, and the computational platforms for efficiently manage this data. First block ends with a review of success examples on real problems in business environment related to big data.

The second block presents the MapReduce model for making use of big databases over Big Data platforms.

The third block is aimed to the design and development of machine learning applications over Big Data platforms. This part also focuses in languages available on Big Data platforms for implementing applications, focusing mainly on Python and R.

PREVIOUS KNOWLEDGE

Relationship to other subjects of the same degree

There are no specified enrollment restrictions with other subjects of the curriculum.

Other requirements

Signal analysis
Exploratory Data Analysis
Machine Learning (I)
Machine Learning (II)
Advanced Data Visualization

OUTCOMES

2221 - M.U. en Ciencia de Datos

- Students should communicate conclusions and underlying knowledge clearly and unambiguously to both specialized and non-specialized audiences.
- Students should demonstrate self-directed learning skills for continued academic growth.
- Students should possess and understand foundational knowledge that enables original thinking and research in the field.
- Be able to assess the need to complete their technical, scientific, language, computer, literary, ethical, social and human education, and to organise their own learning with a high degree of autonomy.
- Ability to access and manage information in different formats for subsequent analysis in order to obtain knowledge from data.
- Ser capaces de acceder a herramientas de información (bibliográficas y de empleo) y utilizarlas apropiadamente.
- Ser capaces de asumir la responsabilidad de su propio desarrollo profesional y de su especialización en uno o más campos de estudio, aplicando los conocimientos adquiridos en la identificación de salidas profesionales y yacimientos de empleo.
- To know and use the different models of data storage and database management systems using programming languages for the definition, query and handling of data.
- Seleccionar, atendiendo a criterios de eficiencia, escalabilidad, tolerancia a fallos y adecuación al entorno de producción el paradigma de datos óptimo en soluciones Big Data. Entender como las técnicas Big Data se utilizan para soportar y realizar la toma
  de decisiones basadas en datos.

LEARNING OUTCOMES

To know the sate of art and success stories on Big Data
To kwow and implement the main stages when solving analysis problems on huge data sets
To know and use the most popular platforms and tools for processing huge data sets
To process and analyse numeric data (signals) and alphanumeric data (text) on Big Data platforms
To know and implement the MapReduce model for huge data set processing on databases.
To applied the introduced models and techniques to Big Data problems

DESCRIPTION OF CONTENTS

1. Introduction to Big Data

tate of art, problems and solutions in Big Data.

2. BigData Platforms

Review of current tools and platforms for huge volumes of data

3. Real cases: Big Data in business.

Current Big Data uses. How companies are using Big Data.

4. BigData Platforms: Hadoop. Map reduce

1.Programming environment and basic MapReduce application examples: development and starting applications.
2.Hadoop architecture, components and software environment.
3.Hadoop and Python

5. Hadoop related projects. Kafka

1.Kafka: Message passing systems
2.Kafka Architecture(Streaming platform).
3.KafkaDevelopment environment.

6. Spark

1. DataFrames, Datasets and SQL.
2. Spark APIs:: PySpark and SparkR.
3. Spark Streaming.
4. API Spark in R: SparkR.
5. Machine Learning Library: MLlib.

7. Processing with big data

Processing of large volumes of data in the cloud, programming models and applications oriented to Big Data (optimization methods, parallelization, data management, automatic machine learning), architecture and computing systems

8. Examples

Google Earth Engine, IBM Watson

WORKLOAD

ACTIVITY	Hours	% To be attended
Theoretical and practical classes	60,00	100
Development of individual work	20,00	0
Study and independent work	12,00	0
Readings supplementary material	3,00	0
Preparation of evaluation activities	12,00	0
Preparing lectures	20,00	0
Preparation of practical classes and problem	13,00	0
Resolution of case studies	10,00	0
TOTAL	**150,00**	

TEACHING METHODOLOGY

The course will combine the theoretical and the practical part, without separating sessions devoted to theory from those devoted to practice. The lessons will be taught in a computer equipped classroom.

In the theoretical part of the classes, the teacher will introduce the concepts and methods Statistics and Optimization, with examples and exercises to be solved by the students.

The practical sessions will be synchronized with the theory. In these sessions, the students will learn by solving problems, exercises and case studies, in order to acquire the skills of this course.

EVALUATION

The educational evaluation of knowledge and skills achieved by the students will be made continuously throughout the course, and will consist in the following blocks of evaluation:

1. Exercises and the class work submitted during the course and / or partial exams: 60% of the final grade.

2. Exams: 20% of the final grade.

Grades earned in paragraph 1 shall be kept in the two examination sittings of the academic year in which they were made, since their evaluation is only possible in the teaching period.

REFERENCES

Basic

- Bowles, M. (2015). Machine Learning in Python: Essential Techniques for Predictive Analysis. Ed. Wiley
- Wes McKinney, W (2012). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Ipython. Ed. O'Reilly Media
- Scalable Big Data Architecture. Bahaaldine Azarmi. Apress, (2016)
- Amit Nandi. (2015).Spark for Python Developers. Ed. Packt Publishing
- Matei Zaharia, Bill Chambers (2017). Spark: The Definitive Guide. Big data processing made simple. Ed. O'Reilly Media.
- Matei Zaharia, Holden Karau, Andy Konwinski, Patrick Wendell (2015). Learning Spark Lightning-Fast Big Data Analysis. Ed. OReilly Media.
- Tom White. (2015) Hadoop: The Definitive Guide. 4th Edition. Storage and Análysis at Internate Scale. O'reilly.
- Zachary Radtka and Donald Miner. (2016) Hadoop with Python. O'reilly
- Kafka: The Definitive Guide: Real-time data and stream processing at scale Neha Narkhede , OReally 2017

ADDENDUM COVID-19

This addendum will only be activated if the health situation requires so and with the prior agreement of the Governing Council

English version is not available