VNIVERSITAT DE VALÈNCIA

Course Guide
36437 Big data

COURSE DATA

Data Subject	
Code	36437
Name	Big data
Cycle	Grade
ECTS Credits	6.0
Academic year	2022 - 2023

Study (s)

Degree	Center	Acad. year	Period
1406 - Degree in Data Science	School of Engineering	4	First term

Subject-matter

Degree	Subject-matter	Character
1406 - Degree in Data Science	12 - Computer Science	Obligatory

Coordination

Name	Department
LIBEROS MASCARELL, ALEJANDRO	242 - Electronic Engineering
RODRIGO BORT, MIGUEL	242 - Electronic Engineering

SUMMARY

The subject Massive Data is a four-month subject, consinting on 6 ECTS credits and is taught during the 1st semester of the fourth year of the Degree in Data Science at the University of Valencia.

This subject aims to consolidate the skills of students in the acquisition, storage and processing of data, as well as to provide the tools for scalability of these techniques to the volume of massive data.

In general, at the end of the course, the students must be able to:
• Understand the concept of massive data or Big Data, as well as its social impact and applications.
• Know how to use the different hardware and software resources for data storage.
• Know how to use hardware and software tools typical of distributed file systems.
• Know how to use basic parallel programming techniques for accessing and processing distributed data systems.
• Know how to develop different advanced algorithms for the exploitation of massive data making use of the distributed computing capabilities of the introduced architectures.

The contact information with the responsible teaching staff is published on the website of the Department of Electronic Engineering (http://www.uv.es/die). The course material (notes, practice scripts, activities, etc.) will be available to students through the Virtual Classroom, the e-learning platform of the University of Valencia (http://aulavirtual.uv.es/).

Teachers will preferably use e-mail to convene evaluation events, comments on this teaching guide and other relevant aspects for the teaching-learning process. Similarly, teachers may introduce continuous assessment tasks or other relevant aspects in the daily development of the classroom, as well as in the Virtual Classroom of the subject.

The theory classes will be taught in Spanish and the practical and laboratory classes as stated in the course file available on the degree website.

PREVIOUS KNOWLEDGE

Relationship to other subjects of the same degree

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

Other requirements

The subject is developed in the last year of the Degree in Data Science, and therefore represents a stage of consolidation and expansion of a large part of the knowledge worked during the degree.
Without having previous requisites for the enrollment of the subject, it is recommended to have previously taken the subjects of Networks and Security, Parallel Programming and Data Storage Infrastructure. It is also recommended to have previously worked with the programming languages studied during the degree, s

COMPETENCES (RD 1393/2007) // LEARNING OUTCOMES (RD 822/2021)

1406 - Degree in Data Science

- (CG02) Ability to solve problems with initiative and creativity and to communicate and transmit knowledge, abilities and skills, which should include the ethical and professional responsibility of the activity of a data scientist.
- (CT03) Ability to defend your own work with rigor and arguments and to expose it in an adequate and accurate way with the use of the necessary means.
- (CT05) Ability to evaluate the advantages and disadvantages of different methodological and / or technological alternatives in different fields of application.
- (CE02) To methodologically know and apply the programming techniques and the algorithms necessary for the efficient processing of information and the computer resolution of problems that use large volumes of data.

- (CE04) 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.
- (CE08) Ability to understand, select and use the infrastructure and the techniques used to handle mass data, according to criteria of efficiency, scalability, security, error tolerance and adaptation to the production environment.
- (CE11) Ability to design and implement data acquisition, its integration, transformation, selection, verification of its quality and veracity from different sources, taking into account its character, heterogeneity and variability.
- (CB2) Students must be able to apply their knowledge to their work or vocation in a professional manner and have acquired the competences required for the preparation and defence of arguments and for problem solving in their field of study.
- (CB5) Students must have developed the learning skills needed to undertake further study with a high degree of autonomy.

LEARNING OUTCOMES (RD 1393/2007) // NO CONTENT (RD 822/2021)

The general objectives of the subject described in the summary together with the general, basic, transversal and specific competences, result in the following learning outcomes (RA).

Identify and describe the storage and processing requirements of big data systems.
RA 1- Know how to use the services and tools offered by operating systems for data storage (CG02, CB5, CE04).
RA2- Know how to use the different levels of local storage, from physical storage devices to file systems (CG03, CB2, CE08, CE11).
RA4- Know the basic structure of physical support used in "Big Data" (CG03, CB2, CE08, CE11, CT05).
RA9- Identify and describe the storage and processing requirements of massive data systems (CG03, CB2, CT05, CE04).

Describe and apply the new parallel and distributed processing paradigms required in big data systems.
RA5- Know the layer structure of computer networks, as well as the main protocols and services used on the Internet and in data processing (CG03, CB5, CE04).
RA6- Know how to use the physical and virtual devices necessary to create and maintain networks of processors (CG02, CB5, CT03, CE02).
RA10- Describe and apply the new parallel and distributed processing paradigms necessary in massive data systems (CG03, CB2, CT03, CE02).

Know the architecture and manage the resources of distributed file systems.
RA3- Know the common techniques of virtualization of resources and know how to select and use them (CG03, CB2, CT05, CE11).
RA7- Know the risks derived from obtaining, processing, storing and exchanging data (CG02, CB2, CT03, CE04).
RA11- Know the architecture and manage the resources of distributed file systems (CG02, CB2, CT03, CE08).

Know and apply different paradigms on architecture, data flow and the big data programming model.

RA8- Select and apply the technical measures that allow maintaining the security of the data collection, processing, storage and exchange systems (CG02, CB2, CT03, CE08).
RA12- Know and apply different paradigms on architecture, data flow and the massive data programming model (CE02, CT03, CB5, CG02).
RA16- Evaluate the performance and scalability of a parallel processing system, establishing and applying the metrics for comparison (CG03, CB2, CT03, CE02).

Know and use the services offered by cloud computing systems for massive data processing.
RA13- Know and use the services offered by cloud computing systems for massive data processing (CG02, CB2, CT05, CE02).
RA14- Identify and describe the architectures of parallel and distributed computers (CE04, CT03, CB2, CG03).
RA15- Know and apply the parallel and distributed programming paradigms, the related programming models and the standards for the development of high performance systems (CG03, CB2, CT03, CE02).
RA17- Design and develop distributed algorithms that exploit the parallelism capabilities of parallel and distributed computing infrastructures (CE02, CT03, CB5, CG02).
RA18- Design and develop programs that efficiently use multiprocessors and distributed architectures for data processing (CE02, CT03, CB5, CG02).

DESCRIPTION OF CONTENTS

1. Introduction to massive data

In this unit the concept of massive data or Big Data will be defined, as well as the dependence of this definition on the available / necessary hardware resources. The impact of this technology in recent years will also be discussed. (RA4)

2. Data storage

This unit will delve into data storage from the point of view of software and hardware resources (physical and virtual) that allow the definition, query, maintenance and manipulation of massive data volumes. The HDFS (Hadoop Distributed File System) data storage model will be introduced and worked with. (RA: 1-4).

3. Distributed data

This unit will work on hardware and software tools for the generation, maintenance, access and treatment of databases distributed in interconnected networks. Special attention will be paid to the protocols that ensure the integrity of the data in distributed file systems and to the virtualization of networks or cloud computing systems, among others. It will work with the HDFS protocol as well as distributed storage systems on cloud platforms such as GoogleCloud (RA: 4-9, 11, 13).

4. Distributed programming

In this unit, the basic techniques for the treatment and processing of data in a distributed / parallel way in networks of processors will be introduced. Special attention will be paid to the MapReduce programming model and the Apache Spark environment. (RA: 5-6, 8-15).

5. Applications and algorithms for distributed programming.

In this unit, students' skills in exploiting big data will be consolidated through the design and evaluation of high-level distributed programming algorithms that exploit the parallelism capabilities of parallel computing infrastructures. We will work on learning Apache Spark on distributed databases with HDFS on cloud platforms such as GoogleCloud (RA: 10-18).

WORKLOAD

ACTIVITY	Hours	% To be attended
Theory classes	34,00	100
Laboratory practices	20,00	100
Classroom practices	6,00	100
Development of group work	20,00	0
Study and independent work	24,00	0
Readings supplementary material	6,00	0
Preparation of evaluation activities	24,00	0
Preparing lectures	6,00	0
Preparation of practical classes and problem	10,00	0
TOTAL	**150,00**	

TEACHING METHODOLOGY

Face-to-face / synchronous classes
In the classes the contents of the subject will be developed using an expository methodology. Student participation will also be encouraged through different means: online questionnaires, open questions or any other tool may be used to assess the degree of consolidation of the subject.
This expository / interactive methodology will alternate with classroom practices in which practical cases will be solved or application cases will be studied, among others.
Most relevant competences: CE02, CE04, CE08, CE11, CT01, CT02, CT03, CB1, CB2, CB3, CB4, CB5, CG01, CG02, CG04, CG05, CG07.

Laboratory practices
The contents worked in class will be consolidated with different compulsory attendance laboratory practices where different practical exercises related to the subject will be proposed. The results of these practices will be delivered in a staggered manner throughout the course for evaluation, according to the indications of the teaching staff.

Most relevant competences: CE02, CE04, CE08, CE11, CT01, CT02, CT03, CT05, CB2, CB4, CG02, CG03, CG04, CG05, CG07.

Preparation of practical and theoretical work
In addition to the tasks described, the students must carry out tasks in a remote manner associated with the preparation of practices and the preparation of reports and reports in which special attention must be paid to the description with rigor and with critical capacity of the different methodologies used. and decisions made (CT03, CT05). In addition, an individual study of the subject will be essential both continuously and for the different evaluation acts. During the course, it is also possible to request the consultation of publications or attendance at talks of interest related to the subject.
Most relevant competences: CE02, CE04, CE08, CE11, CT03, CT05, CB2, CB3, CB4, CG02, CG03, CG04, CG05, CG07.

Individual teaching
Throughout the course, the teaching staff will answer questions both through e-mail, as well as through tutorials in face-to-face format or by teleconference. The form of contact with the teaching staff will be the email as indicated in Virtual Classroom. In addition, the contact information with the responsible teaching staff is published on the website of the Department of Electronic Engineering (http://www.uv.es/die).

E-learning and communication platform
The course material (notes, practice scripts, activities, etc.) will be available to students through the Virtual Classroom, the e-learning platform of the University of Valencia (http://aulavirtual.uv.es/).

Teachers will preferably use e-mail to convene evaluation events, comments on this teaching guide and other relevant aspects for the teaching-learning process. Similarly, teachers may introduce continuous assessment tasks or other relevant aspects in the daily development of the classroom, as well as in the Virtual Classroom of the subject.

EVALUATION

Regarding the evaluation, different dimensions of the teaching-learning process will be taken into account. In the first place, the evaluation system responds to the different competences, learning outcomes and contents to be worked on during the course. Second, both group activities and individual work will be balanced. Finally, the evaluation is proposed as formative, that is, comments will be provided that favor the correction of aspects to be improved detected during the course, either in the daily interaction between students and teachers, through comments in the Virtual Classroom or in sessions of revision.

Both in first and second call, the final grade (NF) responds to the different evaluation systems (SE) according to the following expression:

$$NF = SE1 \cdot 0.5 + SE2 \cdot 0.3 + SE3 \cdot 0.2$$

SE1: Individual objective test. (Evaluation instruments: Ex1 / Ex2)
SE2: Evaluation of practices. (Assessment instruments: Labs / LabEx)
SE3: Continuous evaluation. (Evaluation instruments: EjP / TC / others)

In any case: (1) if the score SE1 has a value lower than 5/10 the NF will be equal to SE1; (2) if the note SE2 has a value less than 5/10, NF will be equal to SE2; (3) NF must be higher in first or second call than 5/10 to pass the subject.

The different evaluation instruments are described below:

Ex1 / Ex2: Individual exam (SE1). It may contain both brief questions, as well as the development of theoretical-practical questions, problems, case studies, etc. You can be asked about any aspect worked on during the course, new problems related to the subject may also appear, as this is considered a useful methodology to assess the consolidation of competencies and content. This test will be carried out according to the school's exam calendar, Ex1 corresponds to the first call and Ex2 to the second. Participation in Ex2 will be mandatory as long as the subject is not passed in the first call, otherwise the grade in the second call will be Not Presented, any exception in this regard must be authorized by the teaching staff.
Most relevant competences: CE02, CE04, CE08, CE11, CT03, CT05, CB2, CB3, CB4, CG02, CG07.

Labs: Laboratories (SE2). During the course, and preferably in teams of two people, laboratory practices will be carried out. Both the development in the classroom and the reports associated with each practice will be evaluated, which must discuss the procedures used rigorously and, where appropriate, answer the questions posed. In addition, preparation tasks for the practice may be requested, which may be equally assessable up to a maximum of one third of the mark corresponding to each practice. To have a grade associated with each practice, attendance will be mandatory.
In the first call, SE2 will be determined by the continuous evaluation of the laboratories through the methods described above.
Copying in any of these activities will be strictly penalized and all laboratory evaluation notes may be canceled. Unjustified and repeated non-attendance at laboratory sessions will result in a laboratory grade of 0 on the first call.
Most relevant competences: CE02, CE04, CE08, CE11, CT01, CT02, CT03, CT05, CB2, CB4, CG02, CG03, CG04, CG05, CG07.

LabEx: Laboratory exam (SE2). Students who do not pass the minimum grade in SE2 in the first call through the method described above, must take a laboratory exam. In this exam, their performance with the tools used during the course and their ability to interpret results, among other skills associated with laboratories, will be evaluated. This exam will be carried out on second call and according to the official calendar.
Most relevant competences: CE02, CE04, CE08, CE11, CT03, CT05, CB2, CB4, CG02, CG03, CG07.

ExP: Individual partial exercise (SE3). During the course and during class hours, a test will be carried out to evaluate the consolidation of content and competences, as well as to give students the opportunity to face exercises similar to those that can be found in Ex1 / Ex2. The contents applicable to said test, as well as the following rules to follow and the date will be communicated during the course. In no case will this test eliminate matter facing Ex1 / Ex2. The weight of this exercise will be 10% of the final grade.
Most relevant competences: CE04, CE08, CE11, CT03, CT05, CB2, CB3, CB4, CG02, CG07.

TC: Collaborative tasks (SE3). The continuous evaluation will be completed through different collaborative tasks that can be proposed both for its realization in the classroom and in a non-face-to-face way. These tasks will be carried out as a team, among others the coordination between different members will be worked s of a team, the discussion to reach consensus solutions, the analysis of relevant

bibliography, etc. Peer and continuous assessment techniques may be used to differentiate the grades of different members of a team. The weight of this section will be 10% of the final grade.
Activities or reports submitted after the deadline will not be taken into account, and will not be recovered.
Most relevant competences: CE02, CE04, CE08, CE11, CT01, CT02, CT03, CB1, CB2, CB3, CB4, CB5, CG01, CG02, CG04, CG05, CG07.

Others (SE3): In any case, the teaching staff reserves the possibility of adding other continuous assessment methods such as: daily observation, attendance and participation controls... which may replace / complement the TC section.

In the second call, and following the minimum grades indicated above, the final grade can be calculated as: NF = SE1·0,6+SE2·0,4

In any case, the evaluation system will be governed by what is established in the Evaluation and Qualification Regulations of the University of Valencia for Bachelor's and Master's degrees.

REFERENCES

Basic

- Large Scale Machine Learning with Python, B. Sjardin, A. Boscheti, L. Massaron.
- Big Data: Principles and Best Practices. N. Marz, J. Warren
- Spark, the definitive guide, B. Chambers, M. Zaharia

Additional

- Learning Spark, H. Karau, A. Konwinski, P. Wendell, M. Zaharia
- Learning PySpark, T. Drabas, D. Lee
- PySpark Cookbook, D. Lee, T. Drabas