Universitat Autònoma de Barcelona

2023/2024

Development of Big Data Applications

Code: 104358
ECTS Credits: 6

Degree	Type	Year	Semester
2503758 Data Engineering	OB	3	2

Contact

Name: Antonio Miguel Espinosa Morales

Email: antoniomiguel.espinosa@uab.cat

Teaching groups languages

You can check it through this link. To consult the language you will need to enter the CODE of the subject. Please note that this information is provisional until 30 November 2023.

Teachers

Remo Lucio Suppi Boldrito

External teachers

Ramon Grau

Prerequisites

Although there are no formally established prerequisites and the subject will provide students with the means to acquire the knowledge described in the syllabus. It is advisable: a good knowledge of programming, of the computer structure, of the operating system at the level of user programmer and of database systems

Objectives and Contextualisation

The objectives of the subject are the study of the main concpets of data intensive applications regarding their reliability, scalabilty and performance.

The subject will introduce data distributed systems and paradigms of data processing, programming models for batch, in-memory and streaming processing, architecture of applications like lambda. It will cover the integrity, accesibility, reliability, consistency and security of large data processing. Also the integration of application designs and infrastructure.

Competences

- Conceive, design and implement efficient and secure data storage systems.
- Conceive, design and implement efficient applications for the analysis and management of big data.
- Plan and manage the available time and resources.
- Students must be capable of collecting and interpreting relevant data (usually within their area of study) in order to make statements that reflect social, scientific or ethical relevant issues.

Learning Outcomes

1. Apply techniques to automate applications' response to dynamic situations (reliability, scalability, emergencies, etc.).
2. Develop applications that can process data on a large scale using batch and streaming paradigms.
3. Plan and manage the available time and resources.
4. Students must be capable of collecting and interpreting relevant data (usually within their area of study) in order to make statements that reflect social, scientific or ethical relevant issues.

Content

1. Introduction to massive data applications
2. Main concepts of data management in massive data environments: reliability, scalability and maintainability. Data models and query languages.
3. Large volume data management. Data warehousing. Main principles of Data Warehousing systems, business intelligence, multidimensional modeling, OLAP operators, ETL processes
4. Introduction to in-memory databases with Redis
5. Large volume data management with Apache Spark tools. Introduction to Spark Dataframes and MLib

Methodology

During the development of the subject we may differentiate three types of teaching activities:

Theoretical classes: general description of the theoretical part of each program topic. The typical structure of a theoretical lesson will be the following: first, an introduction will be made where the objectives of the lesson and the contents to be discussed will be briefly presented. Next, the contents of the lesson will be discussed, including narrative expositions, formal developments that provide the theoretical foundations, including examples that illustrate the application of the discussed contents. Finally, the professor will present the conclusions of the lesson. Throughout the course there will be continuous assessments of groups of topics.

Laboratory classes: The practical part of the theoretical topics will be completed with sessions in the laboratory, where the student will develop a series of programs and should try to solve a specific problem that will be received at the beginning of the semester. Some of these exercises must be delivered on the specified dates. The lab sessions will be developed in groups of two or three students. The subject includes a list of sessions in the laboratory, lasting 2 hours each, where the student will carry out the exercises. The lab report will be delivered in the virtual campus to be evaluated.

Case studies: during the final sessions of the subject, a list of practical cases will be presented to the students. These cases will contain challenges to solve with data sets and business objectives. Students will work in groups to describe a list of conclusions of their work in an oral presentation.

This approach to work is oriented to promote an active learning and developing competencies of organization and planning skills, oral and written communication, teamwork and critical reasoning. The quality of the exercises carried out, their presentation and their functioning will be especially valued.

Annotation: Within the schedule set by the centre or degree programme, 15 minutes of one class will be reserved for students to evaluate their lecturers and their courses or modules through questionnaires.

Activities

Title	Hours	ECTS	Learning Outcomes
Type: Directed			
Exercises	9	0.36	1, 3, 4
Labs	12	0.48	2, 3, 4
Theory	20	0.8	3, 4
Type: Autonomous			
Autonomous study	30	1.2	3, 4
Exercise preparation	20	0.8	1, 3, 4
Lab preparation	32	1.28	2, 3, 4

Assessment

The objective of the assessment process is to verify that the student has achieved the knowledge and skills defined in the objectives of the subject, as well as the associated competences.

These types of activities will be assessed independently, and the weighted sum of them will give the final grade.

Theory (T)

Solution of the laboratory practices (PL)

Completion of a practical case study (PA).

The part of Theory (T) will be evaluated with two individual written exams throughout the course. The final grade of Theory will come out of the weighted sum of the two exams (0.5 * Exam 1 + 0.5 * Exam 2). There will be a second chance to recover this part on the day we have assigned in the June exams week. The parts that have not been passed in the partial theory exams can be recovered. The minimum grade for passing this part is >= 4.5. There is a minimum grade of 4 for each partial evaluation to allow the calculation of the final theory mark.

The part of Laboratory exercises (PL) will be evaluated by group. There are two planned deliveries. The final grade will come from the weighted sum of the deliveries. To pass the PL the minimum mark will have to be >= 4.5. There is only one opportunity to pass this part (this part can't be recovered). Labs delivery is compulsory to be evaluated.

The practical case study (PA) will consist on a presentation on the topic chosen. Students will evaluate their peers and the final mark will include the revision from the subject teachers. The value of this part is 10% of the final mark and given their nature and objective are not recoverable.

The final grade of the subject will be the weighted sum of the grades of each of the four activities: 60% of Theory, 10% Resolutionof individual practical exercises and 30% ofResolution of laboratory exercises. The result will have to be >= 5.

In case a student does not pass the subject due to not reaching the minimum score in any of the mandatory parts (Theory or Laboratory exercises), even though the weighted average is equal or superior to 5, the final grade of the subject will be 4.5.

In the event that the average does not reach 5, the official grade will be the average mark obtained numerically.

If the student delivers any activity, it is understood that he/she is participating in the subject and will be evaluated. If you do not deliver any activity, then it can be considered Non-evaluable.

Granting an honorific matriculation qualification is a decision of the faculty responsible for the subject. The regulations of the UAB indicate that MH can only be awarded to students who have obtained a final grade of 9.00 or more. It can be granted up to 5% of MH of the total number of students enrolled.

The dates of continuous evaluation and assignment delivery will be published on the virtual campus and may be subject to possible changes to adapt to possible incidents; the virtual campus will always inform about these changes since it is understood that the CV is the usual mechanism for exchanging information between professors and students.

For each assessment activity, a place, date and time of revision will be indicated in which the student will be able to review the activity with the professor. In this context, claims can be made about the activity grade, which will be evaluated by the professors responsible for the subject. If the student does not submit to this review, this activity will not be reviewed later.

Note about plagiarism:

Without prejudice to other disciplinary measures deemed appropriate, and in accordance with the current academic regulations, irregularities committed by a student who may leadto a variation of the qualification in an assessable activity will be graded with zero (0). Assessment activitiesqualified in this way and by this procedure will not be recoverable. If it is necessary to pass any of these assessment activities to pass the subject, this subject will be suspended directly, without opportunity to recover it in the same course. These irregularities include, among others:

the total or partial copy of a lab exercise, report, or any other evaluation activity;

let another student to copy;

present a group work not done entirely by the members of the group (applied to all members and not only to those who have not worked);

present as own materials prepared by a third party, even if they are translations or adaptations, and generally works with non-original and exclusive elements of the student;

have communication devices (such as mobile phones, smart watches, pens with camera, etc.) accessible during theoretical-practical assessment tests (individual exams);

talk with classmates during the individual theoretical-practice tests (exams);

copy or attempt to copy from other students during the theoretical-practical assessment tests (exams);

use or attempt to use written material related to the subject during the theoretical-practical evaluation tests (exams), when these have not been explicitly allowed.

If you do not pass the subject due to the fact that some of the evaluation activities do not reach the minimum grade required, the numerical official grade will be the lowest value between 4.5 and the weighted average of the grades. With the exceptions that the "Non-Appraising" qualification will be awarded to students who do not

participate in any of the assessment activities, and that the numerical official grade will be the lowest value between 3.0 and the average Weighted grades in case the student has committed irregularities in an evaluation act (and therefore thesubject cannot be approved by compensation). In future editions of this subject, the student who has committed irregularities in an evaluation act will not be validated any of the assessment activities carried out.

In summary: copy, let copy or plagiarize (or attempt) in any of the assessment activities will lead to a FAIL, not compensable and without validations of parts of the subject in subsequent courses.

Assessment Activities

Title	Weighting	Hours	ECTS	Learning Outcomes
Case studies	10%	5	0.2	1, 3, 4
Individual exam part 1	30%	2	0.08	1, 3, 4
Individual exam part 2	30%	2	0.08	1, 3, 4
Lab delivery	30%	18	0.72	2, 3, 4

Bibliography

Designing Data intensive applications - Martin Kleppmann, O'Reilly, 2017

The Data warehouse ETL toolkit - Ralph Kimball, Joe Caserta. Wiley, 2004

Spark, the definitive guide, Big data processing made simple. Bill Chambers and Matei Zaharia, O'Reilly, 2018

Learning Spark - Lightning fast data analysis - Holden Karau, Andi Konwinski, Patrick Wendell, Matei Zaharia, O'Reilly, 2015

Beginning Scala - Layka, Vishal. Apress; 2nd ed. 2015.

Redis in Action - Josiah L. Carlson. Manning, 2013.

Software

Talend Open Studio for Data Integration

Redis

Apache Spark

Jupyter Notebook

Ubuntu Linux