

Degree programme	Type	Course
Modelling for Science and Engineering	OP	1

Contact lecturer

Name : Eduardo Cesar Cabrera Flores

Email : eduardocesar.cabrera@uab.cat

Teaching staff

Daniel Franco Puentes

Group languages

You can consult this information at the [end](#) of the document.

Prerequisites

It is recommended to have a basic knowledge of programming languages like Python or any another high-level programming language and basic skills of any Linux distribution.

Objectives

The objectives of the module:

- Solve data analysis problems with open source tools
- Understand tool data management limitations and learn criteria to select suitable tools for a specific problem
- Learn data query methodologies related to each technology
- Use Cloud Computing providers to solve data analysis problems
- Apply a data analysis methodology to solve practical problems

By the end of the lectures and practical labs students should have enough knowledge to understand the requirements of typical large data analysis problems in industrial and academic contexts. They should be able to pick some combination of tools and design a solution for a given large data analysis problem. This subject is oriented to develop data problem solving skills. Languages, tools and techniques are described in a data analysis context and students will solve a list of data problems applying the technology described at every chapter.

Learning outcomes

- CA24 (Apply the computational tools for large database analysis to solve problems in the industrial or research field.) Apply the computational tools for large database analysis to solve problems in the industrial or research field.
- CA25 (Integrate computational tools for large database analysis in multidisciplinary work environments.) Integrate computational tools for large database analysis in multidisciplinary work environments.
- CA26 (Work in multidisciplinary teams with the aim of developing projects where large database analysis techniques are applied.) Work in multidisciplinary teams with the aim of developing projects where large database analysis techniques are applied.
- KA19 (Characterise the structure and performance of the different architectures and structures of database organisation and/or management.) Characterise the structure and performance of the different architectures and structures of database organisation and/or management.
- KA20 (Describe the cloud computing tools used for data analysis, evaluating their advantages and limitations.) Describe the cloud computing tools used for data analysis, evaluating their advantages and limitations.
- SA24 (Use specific software to solve data processing problems in distributed systems.) Use specific software to solve data processing problems in distributed systems.
- SA25 (Develop computer applications aimed at modelling a specific process using distributed systems, also evaluating their computational performance.) Develop computer applications aimed at modelling a specific process using distributed systems, also evaluating their computational performance.
- SA26 (Interpret the reports and results of the analysis of a specific database.) Interpret the reports and results of the analysis of a specific database.

Contents

T1: Introduction to Distributed Systems and large data processing systems (2 hours)

T2: Cloud computing (2 hours)

- Introduction to cloud computing
- Data analysis with a cloud computing provider: AWS / Azure

T3: Cluster and supercomputer infrastructures (14 hours)

- Principles of job execution under batch queue systems (SLURM).
- Advanced control of jobs: array jobs, dependencies, process binding, heterogeneous resources (GPUs)
- Creation of large jobs and workflows
- Virtualization and environments

T4: Cloud Networking and Virtual Private Clouds (8 hours)

- Intro to VPC
- Build our VPC and launch a web server tutorial
- VPC Lab

T5: Fault tolerance systems (4 hours)

- Availability zones
- Load balancing and autoscaling tutorial
- ELB Lab

T6: Database Cloud project: relational and DynamoDB implementations (8 hours)

- Intro to RDS and DynamoDB
- Build a database server tutorial
- Distributed database Lab

T7: Serverless services and Lambda (2 hours)

- Intro to Lambda services and serverless computing
- Lambda tutorial
- Lambda Lab

Learning activities and methodology

Title	Hours	ECTS	Learning outcomes
Laboratory	24	0.96	CA24, CA25, SA24
Lectures	38	1.52	KA19, KA20, SA25
Practical exercise development	62	2.48	CA26, SA26

The development of the subject will be based fundamentally on the concept of "learning by doing". There will be initial theoretical sessions for each topic, in which the teacher will present the key concepts and students will be provided with complementary study materials (books, online teaching resources, articles and other technical documentation). With this information, students will work in practical sessions solving exercises and small projects individually or in groups of two people. Students will prepare written reports on the practical work carried out in each topic.

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.

Assessment

Continuous assessment activities

Title	Weight	Hours	ECTS	Learning outcomes
Infrastructure lab	30%	6	0.24	CA24, CA26, SA25, SA26
Lambda Lab	10%	4	0.16	CA26, KA20, SA25, SA26
RDS Lab	20%	6	0.24	CA25, CA26, SA25, SA26
ELB Lab	20%	6	0.24	CA26, SA24, SA25, SA26
VPC Lab	20%	4	0.16	CA26, KA19, SA25, SA26

Evaluation will come out from the combination of the work developed in the laboratory sessions and the submission of the corresponding reports.

In this module, the use of Artificial Intelligence (AI) technologies is permitted exclusively for support tasks, such as literature or information searches, text correction or translations, in the specific activities that will be determined in due course. Students must clearly identify which parts have been generated using this technology, specify the tools used, and include a critical reflection on how they have influenced the process and the final outcome of the activity. A lack of transparency regarding the use of AI in this assessed activity will be considered an academic integrity violation and may result in a partial or full deduction of the activity's mark, or more severe sanctions in more serious cases.

Bibliography

Tanenbaum, Andrew S., Steen, Maarten van. "Distributed systems: principles and paradigms", Prentice Hall, 2nd edition, 2006.

Martin Kleppmann. "Designing Data-Intensive Applications". O'Reilly, 2017.

A. Wittig, M. Wittig. "Amazon Web Services in Action", Manning, 2nd Edition, 2018.

Coulouris, George F; J. Dollimore and T. Kindberg, "Distributed Systems: concepts and design", Addison-Wesley, 5th edition, 2012.

Bell, Charles; Kindahl, Mats; Thalmann, Lars. "MySQL High Availability". O'Reilly, 2010. *Available as electronic resource at UAB library*

Bell, Charles, "MySQL Database Service Revealed: Running MySQL as a Service in the Oracle Cloud Infrastructure", Apress, 1st edition, 2023. *Available as electronic resource at UAB library*

Chang, Fay, et al. "Bigtable: A Distributed Storage System for Structured Data", OSDI, 2006

Dewitt, David, and Jim Gray. "Parallel Database Systems: The Future of High Performance Database Processing", Communications of the ACM 35, no. 6 (1992): 85-98

Schwartz, Baron; Zaitsev, Peter; Tkachenko, Vadim; Zawodny, Jeremy D.; Lentz, Arjen; Balling, Derek J. "High Performance MySQL", O'Reilly, 2008.

Seyed M. M. Tahaghoghi and Hugh E. Williams. "Learning MySQL", O'Reilly, 2006

Nathan Haines. "Beginning Ubuntu for Windows and Mac Users". Apress 2015. *Available as electronic resource at UAB library*

William E. Shotts. "The Linux Command Line". Second Internet Edition. 2013. <http://linuxcommand.org/tlcl.php>

Dan C. Marinescu. "Cloud Computing. Theory and Practice". Morgan-Kaufmann. 2018.

R. Buyya, R. N. Calheiros, A. V. Dastjerdi. "Big data. Principles and paradigms". Morgan-Kaufmann. 2016.

Software

In the subject, we are going to use the last version of the following software platforms and tools

-Linux

-SLURM

-Linux development environment

Course groups and languages

The information provided is provisional until November 30. After this date, you will be able to consult the language of each group through this [link](#). To access the information, you will need to enter the course

CODE

Type of teaching	Group	Language	Semester	Shift
(TEm) Theory (master)	1	English	first semester	afternoon
(PLABm) Practical laboratories (master)	1	English	first semester	afternoon