Universitat Autònoma de Barcelona

2021/2022

Distributed Systems

Code: 44212
ECTS Credits: 6

Degree	Type	Year	Semester
4313136 Modelling for Science and Engineering	OT	0	1

The proposed teaching and assessment methodology that appear in the guide may be subject to changes as a result of the restrictions to face-to-face class attendance imposed by the health authorities.

Contact

Name: Antonio Espinosa Morales

Email: AntonioMiguel.Espinosa@uab.cat

Use of Languages

Principal working language: english (eng)

Teachers

Daniel Franco Puntes

Pedro Luis Pons Pons

Prerequisites

It is recommended to have a basic knowledge of programming languages like Python and basic skills of any Linux distribution.

Objectives and Contextualisation

The objectives of the module:

-Solve data analysis problems with open source tools: Linux, Mysql, Spark

-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 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.

Competences

- Analyse and evaluate parallel and distributed computer architectures, and develop and optimise advanced software for these.

- Communicate and justify conclusions clearly and unambiguously to both specialised and non-specialised audiences.
- Continue the learning process, to a large extent autonomously.
- Integrate knowledge and use it to make judgements in complex situations, with incomplete information, while keeping in mind social and ethical responsibilities.
- Solve problems in new or little-known situations within broader (or multidisciplinary) contexts related to the field of study.
- Take part in research projects and working groups in the field of information engineering and high-performance computation.
- Use acquired knowledge as a basis for originality in the application of ideas, often in a research context.

Learning Outcomes

1. Apply the knowledge acquired in the design of distributed storage systems to designing intensive data and computation applications.
2. Apply various techniques for processing and analysing data in order to prepare these analyses in distributed systems.
3. Choose both the distributed platform and the most appropriate language when formulating a solution to a distributed computation problem.
4. Communicate and justify conclusions clearly and unambiguously to both specialised and non-specialised audiences.
5. Continue the learning process, to a large extent autonomously.
6. Integrate knowledge and use it to make judgements in complex situations, with incomplete information, while keeping in mind social and ethical responsibilities.
7. Know the characteristic techniques of data distribution and management and their cost implications in distributed environments.
8. Solve problems in new or little-known situations within broader (or multidisciplinary) contexts related to the field of study.
9. Use acquired knowledge as a basis for originality in the application of ideas, often in a research context.

Content

T1: Introduction to Distributed Systems and large data processing systems (4 hours)

T2: Linux data processing tools and workflow management (10 hours)

- System architecture
- File systems
- Text processing tools
- Linux workflow management

T3: Relational databases and data processing with MySQL (8 hours)

- Relational data model
- Data modelling
- SQL and problem solving using queries

T4: Data parallel processing with Apache distributed tools (10 hours)

- Limitations of the relational data models with large datasets
- Weak consistency models
- Apache tool ecosystem
- Problem solving with Apache Spark

T5: Cloud computing (4 hours)

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

Methodology

The methodology will combine classroom work and problem solving in laboratory sessions. This planned methodology and proposed assessment could be modified depending on restrictions on physical attendance to University classrooms due to health measures.

Virtual classes and labs will take place in a class Teams virtual space where all students will be invited to access. Lab sessions will be scheduled at the beginning of the course and will use the same Teams space for the development of all practical labs. Students will use a local Linux environment: native, using VirtualBox or using a Cloud Computing instance.

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			
Laboratory	24	0.96	2, 1, 4, 8, 5, 9
Lectures	38	1.52	2, 1, 7, 6, 8, 3, 9
Type: Autonomous			
Practical exercise development	62	2.48	2, 1, 8, 5

Assessment

Evaluation will come out from the combination of work developed in the lab sessions and a final exam.

Assessment Activities

Title	Weighting	Hours	ECTS	Learning Outcomes
Cloud Computing lab	10%	4	0.16	2, 7, 8
Exam	30%	2	0.08	2, 1, 4, 9
Linux lab	20%	6	0.24	2, 7, 8, 5
Mysql lab	20%	6	0.24	2, 1, 7, 6, 8, 3, 9
Spark lab	20%	8	0.32	2, 1, 7, 8, 3, 9

Bibliography

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

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

G. Coulouris, J. Dollimore and T. Kinderg, "Distributed Systems. Concepts and design ", Addison-Wesley, 5th edition, 2012.

Bell, Charles; Kindahl, Mats; Thalmann, Lars. "MySQL High Availability". O'Reilly, 2010.

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. "Saied" 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

Petar Zecevic, Marko Bonaci. "Spark in Action". First Edition. Manning. 2017

V. Layka, D. Pollak. "Beginning Scala". Apress. 2015. *Available as electronic resource at UAB library*

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

-VirtualBox

-Ubuntu Linux

-PostgreSQL

-Apache Spark