Universitat Autònoma de Barcelona

2018/2019

Parallel and Distributed Calculation Systems

Code: 43343
ECTS Credits: 6

Degree	Type	Year	Semester
4313136 Modelling for Science and Engineering	OT	0	1
4314660 Computer Engineering	OB	1	1

Contact

Name: Antonio Espinosa Morales

Email: AntonioMiguel.Espinosa@uab.cat

Use of languages

Principal working language: english (eng)

Teachers

Antonio Espinosa Morales

Prerequisites

It is recommended to have basic knowledge in programming languages like Python, Java or Scala and basic Linux management tasks like installing applications

Objectives and Contextualisation

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 the problem.

Competences

Knowledge:

- Analyze and evaluate distributed computer systems, and software development principles.
- Know the current innovative solutions to distributed systems problems, servers and applications.
- Understand the main principles of large data analysis systems and how to solve large data problems with those systems.

Expertise:

- Use and apply a wide range of design techniques, middleware and development tools for solving a large data problem with a distributed tool platform.
- Be able to select both the distributed platform, such as the most suitable tool, for solving problems of data analysis in a distributed computing context.
- Apply the knowledge acquired of distributed storage systems to design data-intensive applications.

Attitude:

- Demonstrate accountability in the management of information and knowledge, and address groups and / or multidisciplinary projects.

- Apply research methods, techniques and specific resources for research in a particular area of expertise.

Skills

Modelling for Science and Engineering

- 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 a wide range of techniques for designing middleware and development tools to tie together the environment and the application.
2. Apply the knowledge acquired in the design of distributed storage systems to designing intensive data and computation applications.
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. Distinguish the parallel computing environments and their implications in terms of performance and cost.
7. Integrate knowledge and use it to make judgements in complex situations, with incomplete information, while keeping in mind social and ethical responsibilities.
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 workflows (12 hours)

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

T3: Relational databases and data processing with MySQL (12 hours)

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

T4: Data parallel processing with Apache distributed tools (12 hours)

- Limitations of the relational data models with large datasets
- Weak consistency models
- Apache tool ecosystem
- Problem solving with Apache tools: Hive, Hadoop, Spark

T5: Cloud computing (4 hours)

- Cloud application models
- Resource models considering cost and usage

Methodology

The methodology will combine classroom work, problem solving in class and collaborative work in the laboratory delivered as practical reports

Activities

Title	Hours	ECTS	Learning outcomes
Type: Directed			
Final examination	2	0.08	2, 3, 7, 8
Lab work	16	0.64	1, 2, 3, 4, 5, 6, 7, 8, 9
Lectures	24	0.96	1, 6, 7
Presentation work	2	0.08	2, 3, 9
Type: Autonomous			
Study and home works	100	4	2, 5, 9

Evaluation

Evaluation will come out from the combination of: (1) evaluation of lab reports, (2) attendance to lectures and participation in class and presentations of group work, and (3) a final exam.

Evaluation activities

Title	Weighting	Hours	ECTS	Learning outcomes
Exam	30%	2	0.08	2, 3, 7, 8
Lab work	45%	2	0.08	2, 3, 4, 5, 6, 7, 8, 9
Presentation work	25%	2	0.08	1, 2, 3, 9

Bibliography

G. Coulouris, J. Dollimore and T. Kinderg, "Sistemas Distribuidos: Conceptos y Diseño", Addison-Wesley, 3a Ed. 2001.

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

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". 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.