CS-449

Systems for data science

Kermarrec Anne-Marie

Cursus	Sem.	Type
Civil & Environmental Engineering		Opt.
Computational science and Engineering	MA2, MA4	Opt.
Data Science	MA2, MA4	Obl.
Data science minor	E	Opt.

Language of teaching	English
Credits	6
Session	Summer
Semester	Spring
Exam	During the semester
Workload	180h
Weeks	14
Hours	**6 weekly**
Courses	2 weekly
Exercises	2 weekly
Project	2 weekly
Number of positions	

Summary

The course covers fundamental principles for understanding and building systems for managing and analyzing large amounts of data.

Content

Big data systems design and implementation :

- *Distributed systems for data science*
- *Data management : locality, accesses, partitioning, replication*
- *Distributed Machine Learning Systems : federated learning/parameter server/decentralized learning*
- *Massively parallel processing operations*

Large-scale storage systems :

- *Data structures : File systems, Key-value stores, DBMS*
- *Consistency models. The CAP theorem. NoSQL and NewSQL systems*
- *Transactions*

Large-scale processing :

- *Parallel processing*
- *Streaming Processing*
- *Online Processing*
- *Graph Processing*

Keywords

Distributed systems, Parallel programming, Large-scale storage systems, Large-scale data management

Learning Prerequisites

Required courses

CS-322: Introduction to database systems

Recommended courses

CS-322 Introduction to database systems
CS-323: Introduction to operating systems

CS-206 Parallelism and concurrency

Important concepts to start the course

- *Algorithms and data structures.*
- *Scala and/or Java programming languages will be used throughout the course. Programming experience in one of these languages is strongly recommended.*
- *Basic knowledge or computer networking and distributed systems*

Learning Outcomes

By the end of the course, the student must be able to:

- Choose systems parameters, data layouts, and application designs for database systems and applications.
- Develop data-parallel analytics programs that make use of modern clusters and cloud offerings to scale up to very large workloads.
- Analyze the trade-offs between various approaches to large-scala data management and analytics, depending on efficiency, scalability, and latency needs
- Choose the most appropriate existing systems architecture and technology for a task

Teaching methods

Lectures, exercisesand practical work

Expected student activities

During the semester, the students are expected to:

- attend the lectures in order to ask questions and interact with the professor,
- attend the exercises session to solve and discuss exercises,
- solve practical homeworks and/or finish a project during the semester,
- take the exams during the semester

Assessment methods

Homeworks, written examinations, project. Continuous control

Supervision

Office hours	Yes
Assistants	Yes
Forum	Yes
Others	Office ours by appointment

Resources

Bibliography

Relevant resources (textbook chapters, articles, and videos) posted on moodle page.