CS-460

Systems for data management and data science

Ailamaki Anastasia, Kermarrec Anne-Marie

Cursus	Sem.	Type
Civil & Environmental Engineering		Opt.
Computational science and Engineering	MA2, MA4	Opt.
Computational science and engineering minor	E	Opt.
Computer and Communication Sciences		Opt.
Computer science minor	E	Opt.
Computer science	MA2, MA4	Obl.
Cybersecurity	MA2, MA4	Obl.
Data Science	MA2, MA4	Obl.
Data science minor	E	Opt.
Digital Humanities	MA2, MA4	Opt.
SC master EPFL	MA2, MA4	Opt.

Language of teaching	English
Credits	8
Session	Summer
Semester	Spring
Exam	Written
Workload	240h
Weeks	14
Hours	**6 weekly**
Courses	2 weekly
Exercises	2 weekly
Lab	2 weekly
Number of positions	

Summary

This is a course for students who want to understand modern large-scale data analysis systems and database systems. The course covers fundamental principles for understanding and building systems for managing and analyzing large amounts of data. It covers a wide range of topics and technologies.

Content

Topics include large-scale data systems design and implementation, and specifically :

- Distributed data management systems
- Data management : locality, accesses, partitioning, replication
- Modern storage hierarchies
- Query optimization, database tuning
- Transaction management
- Data structures : File systems, Key-value stores, DBMS
- Consistency models
- Large-scale data analytics infrastructures
- Parallel Processing
- Data stream and graph processing

Learning Prerequisites

Required courses

- CS-107 Introduction to programming
- CS-214 Software construction
- CS-300 Data-Intensive Systems
- CS-202 Computer systems

or equivalent courses

Important concepts to start the course

- Knowledge of 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 recommanded.
- Basic knowledge or computer networking and distributed systems.

Learning Outcomes

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

- Understand in detail the design big data analytics systems using state-of-the-art infrastructures for horizontal scaling, e.g., Spark
- Implement algorithms and data structures for streaming data analytics
- Understand the advantage and disavantages of different storage models for a given workload, based on the offered optimization enabled by each model and the workload characteristics
- Compare concurrency control algorithms, and algorithms for distributed data management
- Configure systems parameters, data layouts, and application designs for database systems
- Develop data-parallel analytics programs that make us of modern clusters and cloud offerings to scale up to very large workloads
- Analyze the trade-offs between various approaches to large-scale data management and analytics, depending on efficiency, scalability, and latency needs

Teaching methods

Lectures, project, homework, exercises and practical work

Expected student activities

- Attend lectures and participate in class
- Complete a project as per the guidelines posted by the teaching team

Assessment methods

- Project
- Midterm (as needed)
- Final exam

Supervision

Office hours	Yes
Assistants	Yes
Forum	Yes

Resources

Bibliography

J. Hellerstein & M. Stonebraker, Readings in Database Systems, 4th Edition, 2005
R. Ramakrishnan & J. Gehrke: "Database Management Systems", McGraw-Hill, 3rd Edition, 2002.
A. Rajaraman & J. Ullman: "Mining of Massive Datasets", Cambridge Univ. Press, 2011.

Ressources en bibliothèque

- Mining of Massive Datasets / Rajaraman
- Database Management Systems / Ramakrishnan
- Readings in Database Systems / Hellerstein

Moodle Link

- https://go.epfl.ch/CS-460