CS-460  Systems for data management and data science
Ailamaki Anastasia, Kermarrec Anne-Marie

Summary
This course is intended 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, access, 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
Recommended courses
- CS-107 Introduction to programming
- CS-206 Parallelism and concurrency
- CS-322 Introduction to database systems
- CS-323 Introduction to operating systems
- CS-452 Foundations of software

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:
• Understand how to 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
• Decide between different storage models based on the offered optimizations enabled by each model and the expected query workload
• 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 use 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

Ressources en bibliothèque
• Mining of Massive Datasets / Rajaraman
• Readings in Database Systems / Hellerstein
• Database Management Systems / Ramakrishnan

Moodle Link
• https://go.epfl.ch/CS-460