FID3019 Advanced course in Data-Intensive Computing 7.5 credits

Avancerad kurs i data-intensiv databehandling

This is a translation of the Swedish, legally binding, course syllabus.

Establishment

Course syllabus for FID3019 valid from Autumn 2017

Grading scale

G

Education cycle

Third cycle

Specific prerequisites

Enrolled as a doctoral student.

Language of instruction

The language of instruction is specified in the course offering information in the course catalogue.

Intended learning outcomes

The course complements distributed systems courses, with a focus on processing, storing and analyzing massive data. It prepares the students for Ph.D. studies in the area of data-intensive computing systems.
The main objective of this course is to provide the students with a solid foundation for understanding large scale distributed systems used for storing and processing massive data. More specifically after the course is completed the student will be able to:

- Explain the architecture and properties of the computer systems needed to store, search and index large volumes of data.
- Describe the different computational models for processing large data sets for data at rest (batch processing) and data in motion (stream processing).
- Use various computational engines to design and implements nontrivial analytics on massive data.
- Explain the different models for scheduling and resource allocation computational tasks on large computing clusters.
- Elaborate on the tradeoffs when designing efficient algorithms for processing massive data in a distributed computing setting.

Course contents

Topics:

- Distributed file systems
- No SQL databases
- Scalable messaging systems
- Big Data execution engines: Map-Reduce, Spark
- High level queries and interactive processing: Hive and Spark SQL
- Stream processing
- Graph processing
- Scalable machine learning
- Resource management

Disposition

The course is organized as a reading course. Each student chooses a set of papers and for each paper the student will do the following:
* carefully read and analyze the paper.
* orally present the paper's content including methodology and contributions to the other course participants and the course's examiner. The presentation including discussion should take around one hour.
* write a critical review of the paper that covers in particular: summary of contributions, methodology, significance, technical and experimental quality, and quality of presentation.
In addition to presenting four papers each, the students must read some of the papers assigned to the other course participants, attend their presentations, and actively contribute to the discussion of their papers.

Course literature

Latest papers in the area of Data intensive Computing from high-quality international venues.

Examination

Based on recommendation from KTH’s coordinator for disabilities, the examiner will decide how to adapt an examination for students with documented disability.

The examiner may apply another examination format when re-examining individual students.

If the course is discontinued, students may request to be examined during the following two academic years.

P/F

Other requirements for final grade

The course will be assessed with a Pass/Fail grade, based on a successful presentation, the delivery of a scientifically sound report, and the identification of appropriate papers for the reading list. In addition to this, students must attend at least 75% of all seminars.

Ethical approach

- All members of a group are responsible for the group's work.
- In any assessment, every student shall honestly disclose any help received and sources used.
- In an oral assessment, every student shall be able to present and answer questions about the entire assignment and solution.