2IMW10: Data Engineering (5 ECTS)
Quartile 4 2016-2017
Mondays 13:45-15:30 & Wednesday 8:45-10:30

Staff

dr. George Fletcher (responsible lecturer)
email: g.h.l. + "family name" + @tue.nl
web: `http://www.win.tue.nl/~gfletche/`
office: MF 7.063

dr. Nikolay Yakovets (co-lecturer)
email: n. + "family name" + @tue.nl
web: `http://yakovets.ca`
office: MF 7.099

Prior Knowledge

2ID50 – Data modeling and databases (mandatory)

Learning Objectives

Students will (1) know the main characteristics and relevant research results for models of contemporary data intensive systems; (2) understand the practical relevance of these models for engineering data intensive applications; (3) understand the relative advantages and disadvantages of these models and acquire the ability to decide, based on a problem description, which model is best suited to solve this problem; and, (4) be able to quickly master and make practical use of contemporary frameworks and technologies implementing these models.

Major Topics Covered in the Course

We study models of contemporary data intensive systems and their practical use. These models are among: Graph databases, Data warehousing and online analytical processing (ROLAP, MOLAP, etc.), Document databases (NoSQL, JSON stores, etc.), Parallel and distributed data processing (MapReduce, etc.), and Deductive databases (Datalog).

We discuss why these models were introduced, their relative advantages and disadvantages, how to use them in practice, and, at a high level, how they are implemented. Unlike the subject Database Technology (2IMW20) which focuses primarily on systems internals and their efficient implementation at a lower level, the primary goal of this subject is to develop the practical ability to engineer non-trivial data intensive applications based on a solid understanding of the underlying engineering principles. Towards this goal, hands-on practical assignment(s) using contemporary frameworks and technologies are a central component of the course.

Required Textbook

There is no required textbook. Readings will be posted on the course website.

Student Responsibilities and Grading Criteria

- Individual course participation (15%),
- One final exam (25%), and
- One multi-part team project (60%).

Tentative Week-by-Week Course Outline

- 24-30 April. Course introduction, Graph databases
- 1-7 May. Document databases, Parallel and distributed data processing
  - First team report due on Saturday 6 May
- 8-14 May. Team meetings with instructors
- 15-21 May. Work on team projects
- 22-28 May. Data engineering for data sciences, Data warehousing
  - Second team report due on Saturday 27 May
- 29 May – 4 June. Work on team projects
  - Third team report due on Saturday 3 June
- 5-11 June. Team meetings with instructors
- 12-18 June. Deductive databases, Course conclusions & summary
- 19-25 June. Poster Sessions I & II
  - Final team report due by end of Friday 30 June
- Final exam. Written take-home exam. Due by end of Friday 7 July.

Course Policies

- Participation. As this class endeavors to teach professional skills, it is reasonable to ask that students act professionally and treat all course participants with respect. The subject matter of this course deserves discussion; I encourage you to offer your ideas and thoughts to the class and to question the material presented.
- Assignments & project deliverables are due at the time and in the manner specified in the assignment description. Late work will lose 33% of its original point-value for each day late, and once solutions are posted or discussed late submissions will not be accepted.
- Plagiarism and cheating will not be tolerated. University policy will be adhered to in all such cases. You are free to work with others in interpreting assignments, practicing with tools, and inspecting code. However, individual assignments and the exam are to be done individually. Submissions that appear to be plagiarized will trigger an investigation.