Big Data technologies for Data Science

Studiegidsnummer	
Studielast	6
Voertaal	English
Periode(n)	4
Onderwijsinstituut	Informatics Institute
Inlichtingen	Peter Boncz (boncz@cwi.nl)
Onderdeel van	?

Learning goals After this course, students should be able to:

- Work with large amounts of real-world data.
- Formulate feasible data science problems that can be solved using big data analysis pipelines, and understand their complexity and requirements in terms of infrastructure.
- Design a pipeline for big data analysis, from data collection to analytics and visualization, and tune it to answer specific (types of) questions.
- Use the most common Big Data tools (e.g., Hadoop, Spark, graph databases), infrastructure (e.g., cloud computing), and programming paradigms (map-reduce, stream processing, etc.) to implement the designed pipelines.
- Understand performance and efficiency metrics. Evaluate the performance and efficiency of big data solutions. Analyze performance bottlenecks.
- Work in project teams with people from a variety of backgrounds towards solving a (big) data science problem at all its layers (i.e., from raw data to final visualizations of the results).

Content

This course aims to provide an introduction into the main challenges that big data applications pose across all layers of a processing system, from its infrastructure to its performance. The common solutions - from design to implementation - that are being used to tackle these problems will be presented. Specifically, students will be introduced to storage and processing solutions, infrastructure options, performance challenges, systems and tools for data analytics at scale. Finally, the different success metrics to be used for these solutions will be introduced. Additionally, ethical concerns, as well as interaction with traditional data producers and consumers will be discussed. Therefore, upon completing this course, the students should be able to design a big data analysis framework, reason about its infrastructure requirements, and provide prototype implementation using modern tools and technologies.

Set up

- lectures : 2h per week

- labs: 2h per week, assignment-driven.

- seminar: 1h per week for invited "experience talks" - first by experts, then by students

- project: 1h per week (coupled with the seminar) for discussing project progress

Week 1	**Big Data and Cloud Computing**
Week 2	**The MapReduce Framework**
Week 3	**The Hadoop Ecosystem**
Week 4	**Spark and MLLib**
Week 5	**SQL on Big Data**
Week 6	**noSQL Systems**
Week 7	**Data Streams**
Week 8	**Exam.**

teachers:

Hannes Mühleisen, Peter Boncz, CWI

Grading:

- **25% exam - final exam**
- **25% assignments**
- **50% project (teams of students - 2 or 4)**

Techniques/tools/datasets

- Hadoop/Spark/Pig
- Datasets from different sources: statistics bureau, deltares, competitions of data mining, etc.
- Assignments: shorts tasks to be performed on various Big Data technologies via Amazon Web Services
- Project: design + implementation + presentation + short report

Studymaterial (books, literature)

- See the ‘Technical Literature’ sections in the lecture description on www.cwi.nl/~boncz/bigdatacourse