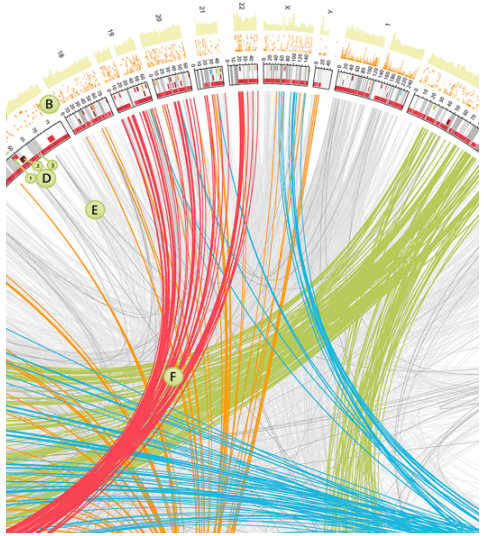
Big Data Analytics — Welcome

Introduction

- You
  - Who is who?
  - Why are you here?
  - What is your experience related to this subject?
    - Programming in Java
    - Data analytics
- Us
  - Who are we?
  - What are our domains of competence?
  - Why are we here?
  - How to contact us
    - Nastaran.Fatemi@heig-vd.ch
    - Marcel.Graf@heig-vd.ch
    - Fatemeh.Borran@heig-vd.ch

Course objectives

- This course presents techniques to manipulate, store and analyze large volumes of data
- The accent of the course is put on two paradigms for the design and implementation of algorithms:
  - Using MapReduce and the Apache Hadoop framework.
  - Using Resilient Distributed Datasets (RDDs) and the Apache Spark framework.
- MapReduce is one of the most well known recent programming models for parallel and distributed computing, used widely in industrial applications dealing with Big Data.
- RDD is emerging as an important programming model taking advantage of functional programming.

Organization

- Lectures
  - Presentation of theoretical concepts
- Labs
  - Gain practical experience with Hadoop, HDFS, MapReduce and Spark and RDD programming
  - Work in a team of two
- Project
  - Work in a team of two on a self-chosen Big Data project

Course structure
(subject to change)

- S01 Introduction to MapReduce (MGF)
- S02 MapReduce advanced topics (MGF)
- S03 MapReduce algorithm design (MGF)
- S04 MapReduce algorithm design 2 (MGF)
- S05 Inverted index (MGF)
- S06 Introduction to Scala collections and their operations (NFI)
- S07 Programming with Spark (NFI)
- S08 Spark lab 1 (NFI)
- S09 Programming with Spark (NFI)
- S10 Spark lab 2 (NFI)
- S11 Project
- S12 Project
- S13 Project
- S14 Project presentations

Evaluation

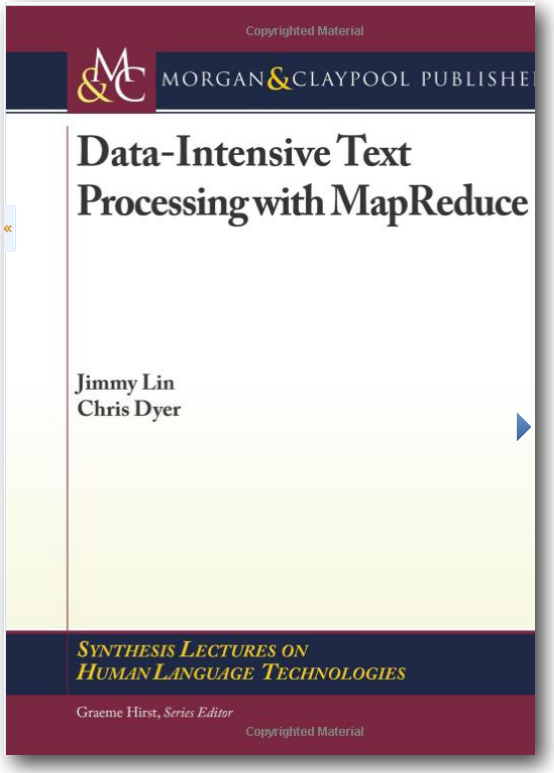
Activity	Coefficient
Labs	25%
Project	25%
Final written exam	50%

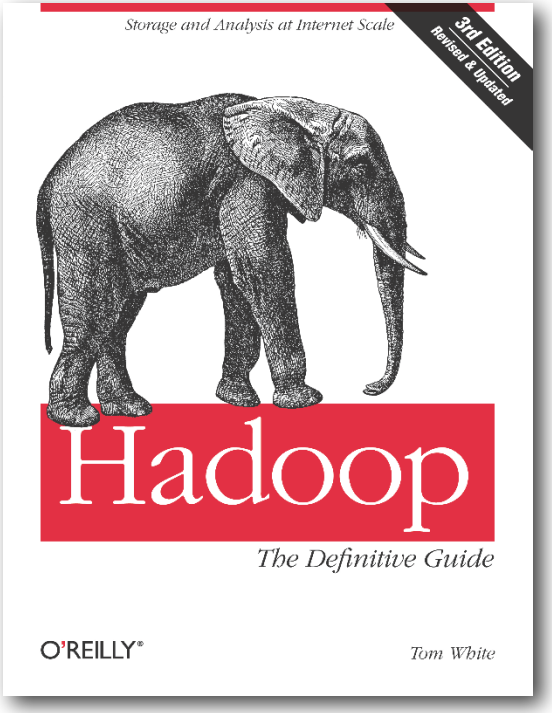
Documentation

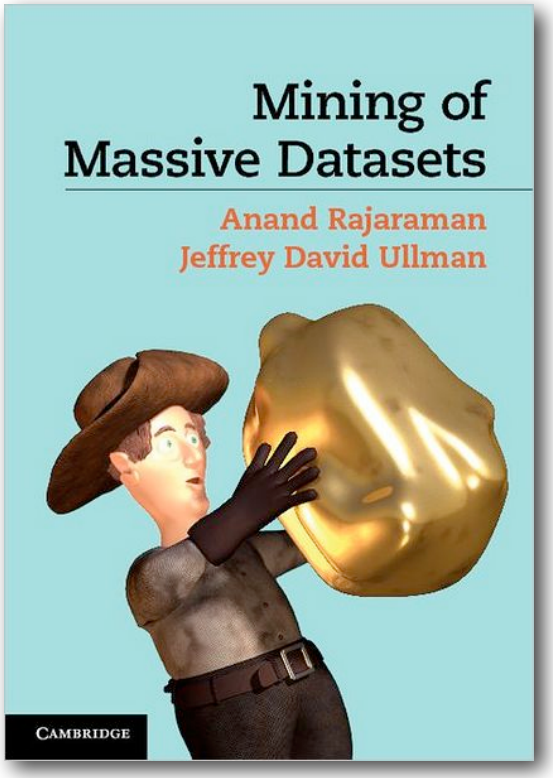
- Lecture slides
- Lab documentation
- Homework
- ... are available on the course web site

http://mse-bda.s3-website-eu-west-1.amazonaws.com

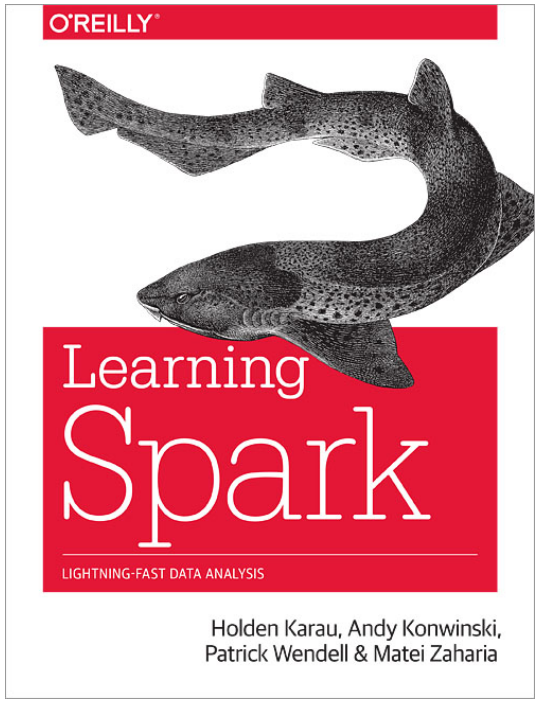
Recommended books

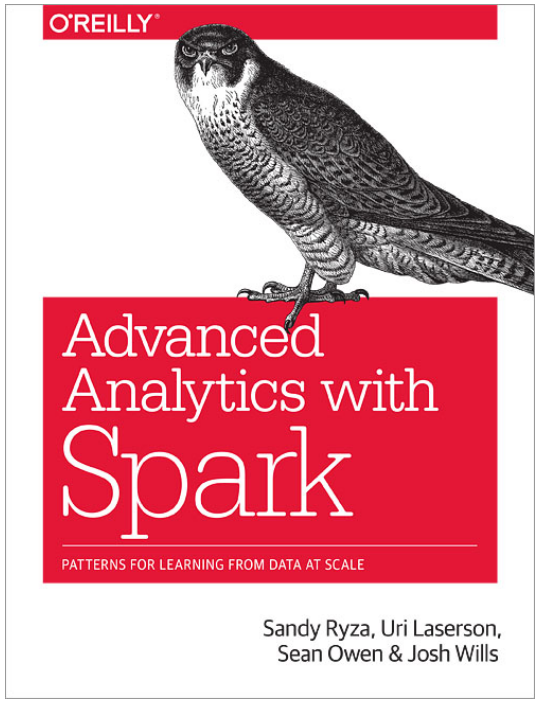


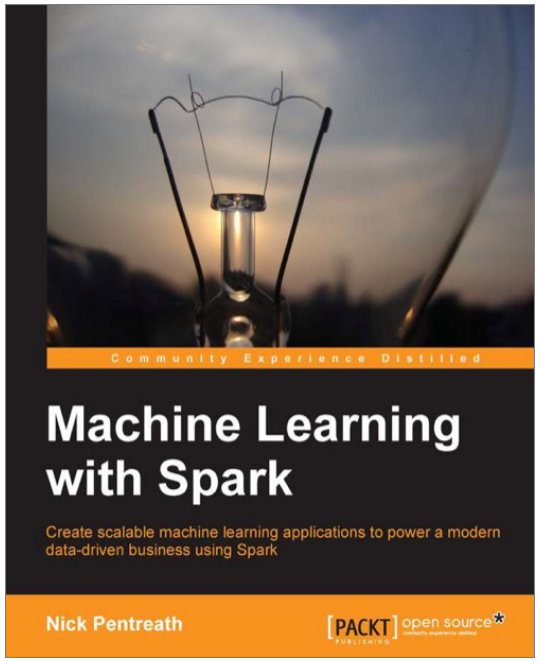




Recommended books







Recommended books and web resources

- Books
  - Jimmy Lin, Chris Dyer — Data-Intensive Text Processing with MapReduce
    Morgan & Claypool
    available at https://github.com/lintool/MapReduceAlgorithms/blob/master/MapReduce-book-final.pdf
  - Tom White — Hadoop, The Definitive Guide
    O'Reilly Media
  - Anand Rajaraman, Jure Leskovec, Jeffrey D. Ullman — Mining of Massive Datasets
    Cambridge University Press
    available at http://infolab.stanford.edu/~ullman/mmds.html
  - Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills — Advanced Analytics with Spark, Patterns for Learning from Data at Scale
    O'Reilly Media
  - Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia — Learning Spark, Lightning-Fast Big Data Analysis
    O'ReillyMedia
  - Nick Pentreath — Machine Learning with Spark
    Packt Publishing
- On the web
  - Stack Overflow : http://stackoverflow.com/ tags: hadoop, mapreduce, apache-spark

Your questions