Data-intensive Scalable Computing Systems

Introduction

Pietro Michiardi

Eurecom

Introduction to the Course

What is this Course About

- **Principles of functional programming**
- **In-depth description of Hadoop MapReduce**
  - Architecture internals
  - Cluster deployments
- **In-depth description of Apache Spark**
  - Architecture internals
- **Relational Algebra and High-Level Languages**
  - Basic operators and their equivalence in MapReduce
  - Apache SparkSQL

What is this Course About

- **Cluster schedulers**
  - Apache YARN, a.k.a. Hadoop v.2
  - Apache Mesos
  - Google Omega
- **Distributed Database Systems**
  - Amazon Dynamo
  - Apache Cassandra
  - Apache HBase
- **Coordination**
  - Apache Zookeeper

Who is this course for?

- **System engineers**
- **Data scientists**
- **Requirements**
  - Good knowledge of Python
  - Familiarity with operating systems concepts, and Linux
  - Good knowledge of git
  - Ideally, familiarity with distributed algorithms

How to make the most of this course?

- **Contribute!**
  - The whole course is open source
  - Pull-request based
  - Contribute to both lecture notes and laboratories
- **Attend classes and the labs**
  - Many discussions in live classes, that are not on the slides
  - Laboratories can be hard for people with little CS background
- **Resources**
  - Lecture notes: `http://michiard.github.io/DISC-CLOUD-COURSE/`

Grading

- **Final exam**
  - 50% of the grade
  - Generally divided in two parts
    - A series of questions
    - One or more problems to solve

- **Laboratory sessions**
  - Mainly Notebooks, some special labs
  - Question answering
  - Heuristic to map credits to grade