Data-intensive Scalable Computing Systems

Introduction

Pietro Michiardi

Eurecom

Introduction and Motivations

What is this Course About

- **The MapReduce Programming Model**
  - Principles of functional programming
- **In-depth description of Hadoop MapReduce v.1**
  - 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
  - Hadoop Pig and PigLatin

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**
  - Familiarity with Java
  - Familiarity with operating systems concepts, and Linux
  - Familiartiy with git
  - Ideally, familiarity with Python and Scala
  - 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/`
  - Laboratories: `https://github.com/michiard/CLOUDS-LAB`

Grading

- **Final exam**
  - 50% of the grade
  - Generally divided in two parts
    - ★ A series of questions
    - ★ One or more problems to solve
  - No coding is required

- **Laboratory sessions**
  - Questions to be answered during the labs
  - Each correct question brings some credits
  - Heuristic to map credits to grade

What is Big Data?

- **Vast repositories of data**
  - The Web
  - Physics
  - Astronomy
  - Finance
- **Volume, Velocity, Variety**
- **It's not the algorithm, it's the data!**
  - More data leads to better accuracy
  - With more data, accuracy of different algorithms converges

What is MapReduce?

- **A programming model**:
  - Inspired by functional programming
  - Parallel computations on massive amounts of data
- **An execution framework**:
  - Designed for large-scale data processing
  - Designed to run on clusters of commodity hardware