Technologies for Big Data with PYTHON

marco milanesio

MS DATA SCIENCE 2019-2020

Who am I?

- PhD [2010]
  - Distributed systems
  - Network measurements and performances
  - Distributed storage
- @Inria
  - Optimization
  - Image processing
  - Spark
  - HPC - cluster computing
- @MDLab
  - Dev-ops
  - Implementation - optimization
  - Virtualization

Who are you?

- Curious, open minded
- Wanting to learn some cool stuff
- Not feared of tackling problems
- Not feared by errors
- (optional) Some coding experience
- (bonus) Some Python experience
- (bonus) "LMGTFY" skills 😎

- Ideal profile:
  - 60% data scientist: exploit the data
  - 30% software developer: you need to code
  - 10% system engineer: because it's fun!

Course Overview

- Introduction

- The Python3 language
- Basic data analysis

- The BigData picture
- The Distributed Computing approach
- The MapReduce framework
- Apache Spark

Course Overview

- First part
  - Syntax, data structures, types
  - Builtins
  - Libraries
  - Data analysis
- Second part
  - What is Big Data (really)
  - Principles of functional programming
  - Distributed applications
  - (optional) Spark MLlib

Course Overview

- 10 lessons
- 40% lectures
- 60% lab sessions
- Evaluation (to be discussed)