INTRODUCTION TO APACHE SPARK

0. COURSE OVERVIEW

WHO AM I?

- Maryan Morel
- PhD student at CMAP / École Polytechnique
- Soon ML engineer at Powder.gg
- Research in ML and Datascience (Python, C++)
- Data engineering (Python, Scala)
- e-mail address: maryan.morel@polytechnique.edu
- Better to use Slack (#advanced_topics_in_databases) than email though :)

OUTLINE

1. Distributed computing and Big Data (04/02)
2. Apache Spark and Resilient Distributed Datasets (RDDs) (10/02)
3. Spark SQL and DataFrames (24/02)
4. Machine learning with Spark MLlib (25/02)

COURSE LOGISTICS

- 18 hours course, class + lab
- Homeworks: Send me your solutions for the next week
- ...except for the last one, for which you have two weeks
- Each homework counts for 1/4 of the final grade
- Penalty of 1 point per day of delay
- Only homework 2 and 3 will be graded for students in apprenticeship.

HANDS-ON LABS

- Hands-on labs and homeworks are based on Jupyter Notebooks
- Notebooks will be released each week from the course homepage

QUESTIONS?