Introduction to Big Data

Lesson 1.0 Course Introduction

Arthur Katossky & Rémi Pépin

Wednesday, March 30, 2022

An intoduction to something *big*

Two revolutions at the same time

- Big data revolution : data science oriented
- Cloud computing : IT oriented

This course is a **introduction** to those two revolutions. So you won't be experts on those fields. But you will learn a lot about computers and IT in general

The two speakers

- Rémi Pepin : teaching assistant in the IT department and former Java developer at Insee
- Arhur Katossky : NLP enthusiast, currently student at Ensae, former teaching assistants in the economy department.

The objectives

1. Understand the basics of computation in the real world, the bottlenecks and how to solve them
2. Understand the basics of cloud computing and how to use AWS
3. Get familiar with big data technologies and the most common paradigm
4. Learn how to use Spark for data exploration on data at rest or streamed data, and how to some basics ML algorithm on big data

How you will be graded

- One graded lab at the end of the coures
- One quick exam after the course

Last warning

The labs are more like tutorials than practical session. Because you will discover new things, labs follow the "To do X you have to use Y" paradigm.

If you think it's too easy, you can explore things by yourself =)

A little game : which is the fastest ?

- US weather data : 1 file per year, 71 years of data, around 6Go when gziped
- Extract the max temperature of each year
- 7 contestants :
  - a classic R script with only R core function
  - a classic python loop (loop through file + read each line one by one)
  - a same but with Cython (python compiled in C)
  - the same code but in C. This code is compiled to machine code
  - The same code but in java. This code is compiled to byte code (need JVM)
  - python but each file are process in parallel (with 12 cores)
  - a awk command in bash to read each line of each file