DATAAI922: Big Data Processing

Louis Jachiet
About the course
Pedagogical Team

Louis Jachiet
Associate Professor
Main teacher

Nedeljko Radulovic
PhD
Teacher Assistant
Objective of the course

Cover various Big Data platforms

Cover various Big Data platforms

- Spark
- Flink
- beam
Objective of the course

Cover various Big Data platforms

Provide knowledge on their construction
Objective of the course

Cover various Big Data platforms

Provide knowledge on their construction

Some practical experience with Hadoop MapReduce and Spark

FileOutputFormat.setOutputPath(job, new Path("out1"));
job.setMapperClass(Map.class);
job.setReducerClass(Reduce.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);
The lab sessions will be graded.
The lab sessions will be graded.

The final will be half lab session and half final exam.
What is Big Data?
"Big data" is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.

– Wikipedia @ 09/09/2019
The three Vs

Figure 1: source Wikipedia
Big Data depends on the context.

Different context will call for different tools!
What we will cover in this course
• Week 1: Course: Introduction & Hadoop
• Week 2: Lab session: Hadoop
• Week 3: Course: Spark RDD
• Week 4: Lab session: Spark
• Week 5: Course: Spark Dataframe
• Week 6: Lab session: Spark Dataframe
• Week 7: Course: Stream processing and Kafka/Lambda architectures
• Week 8: Exam
Course organization

- Lab session take place physically at Telecom but you can also join online
- I will try to make everything available asynchronously online (warn me if something is not recorded!)
- I tried to increase the amount of material available and there might too much exercises but you don’t have to finish them. Only lab sessions are graded!