DATAAI922: Big Data Processing

Louis Jachiet
About the course
<table>
<thead>
<tr>
<th>Date</th>
<th>Location</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>15/09</td>
<td>Amphi 4</td>
<td>Intro &amp; Hadoop</td>
</tr>
<tr>
<td>22/09</td>
<td>Amphi Thévenin</td>
<td>Spark RDD</td>
</tr>
<tr>
<td>29/09</td>
<td>1A207</td>
<td>Lab on Hadoop &amp; Spark RDD</td>
</tr>
<tr>
<td>06/10</td>
<td>Amphi 2</td>
<td>Spark DF</td>
</tr>
<tr>
<td>13/10</td>
<td>1A207</td>
<td>Lab on Spark DF</td>
</tr>
<tr>
<td>20/10</td>
<td>Amphi 2</td>
<td>Streams in Spark &amp; Kafka</td>
</tr>
<tr>
<td>27/10</td>
<td>1A207</td>
<td>Lab on stream &amp; Kafka</td>
</tr>
<tr>
<td>10/11</td>
<td>1C47</td>
<td>Exam</td>
</tr>
</tbody>
</table>
Objective of the course

Cover various Big Data platforms

- Spark
- Flink
- beam
Objective of the course

Cover various Big Data platforms

Provide hindsight on their design
Objective of the course

Cover various Big Data platforms

FileOutputFormat.setOutputPath(job, new Path("out1"));
job.setMapperClass(Map.class);
job.setReducerClass(Reduce.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);

Provide hindsight on their design

Mix theory and practice
The lab sessions will be graded.
The lab sessions will be graded.

The final will be half labs 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 contexts will call for different tools!
Distributed computation
Parallel computing

Characteristics

- one or several computers
- working in very close coordination
- with a very good communication speed (throughput and latency)
- mainly for complex computation

Typical usage:

Any heavy computation, e.g. linear algebra over large data, sat solving, etc.
Distributed computing

Characteristics

- several computers
- working cooperatively with message passing
- relatively good throughput but latency not guaranteed
- mainly for data computation

Typical usage

Computation over large datasets, the typical “embarrassingly parallel” computation.
Federated computing

Characteristics

- several computers
- semi-autonomous
- mainly for uncoordinated parties

Typical usage

Computation between independent parties or “embarrassingly parallel” computations that can be verified.
Knowledge assessment
Using several computers instead of one has the following effect on computation time:

A) it *always* decreases

B) it *might* decrease

C) it *stays* the same

D) it *might* decrease or increase or stay the same
With the most efficient program, using several computers instead of one has the following effect on computation time:

A) it always decreases

B) it might decrease

C) it stays the same

D) it might decrease or increase or stay the same
On a “parallelizable” problem, the speed-up with more computers is:

A) linear

B) sub-linear
With enough computers, Hadoop (and the like such as Spark and Flink) are, in comparison with a manually written programs:

A) *always* faster

B) *sometimes* faster

C) *never* faster
Using several computers instead of one has the following effect on computation time:

A) it always decreases

B) it might decrease

C) it stays the same

D) it might decrease or increase or stay the same