Performance evaluation of Big Data applications

Danilo Ardagna, DEIB

Research goals

- Big data applications require run on many node and include complex software stack
- Research question: given a cluster configuration predict jobs execution time
- Reference technology: Hadoop MapReduce and Tez, Apache Spark

Use of PoliCloud and lessons learned

- Early adopters, we started September 2015

- PRO:
  - Flexibility (free choice of Linux distribution and software versions)
  - Good for benchmark scripts development and testing

- CONS:
  - Network and disks are the performance bottleneck
  - Don’t use for production/final analyses
  - Performance isolation