Calcul distribué sur Spark

Proposé par J.-B. Caillau* M. Dautrey† C. Le Lannou‡

Dans le cas de simulations de grande taille fortement parallélisables, le calcul et l'accès aux données distribués à l'aide de Spark [6] présente de nombreux avantages de plus en plus utilisés pour les applications aux données massives. Du point de vue du stockage et de l'accès aux données, citons notamment la structure de base de *Resilient Distributed Dataset* (RDD). L'objectif de ce projet est de vous permettre de vous familiariser avec ce type d'approche distribuée sur des exemples : un premier cas test sera un calcul de *Value at Risk* (VaR) à l'aide de simulations massives par la méthode Monte-Carlo [2] ; une estimation de type *Kernel Density Estimation* [5] pourra être réalisée en complément. Le développement pourra fait en *Scala* [4] ou *Python*, en s'appuyant sur la librairie de machine learning `MLlib` [3] et sur une plateforme Hadoop/Spark type Hortonworks sur AWS [1] pour les aspects cloud computing.

Références

[1] AWS: Amazon Web Services. `aws.amazon.com`

[2] Horton, G. A. Value-at-Risk: Theory and practice. `value-at-risk.net`

[3] MLlib: Apache Spark's scalable machine learning library. `spark.apache.org/mllib`

[4] Scala: Object-Oriented Meets Functional. `scala-lang.org`

[5] Sheather, S. J. Density estimation. *Statist. Sci.* **19** (2004), no. 4, 588–597.

[6] Spark: A fast and general engine for large-scale data processing. `spark.apache.org`

*UBFC & CNRS/Inria
†Isitix, Perros-Guirec
‡DataLearning Ltd, Londres