# Approche préventive pour une gestion élastique du traitement parallèle et distribué de flux de données

Roland Kotto-Kombi*,Nicolas Lumineau**, Philippe Lamarre*

* Univ Lyon, INSA de Lyon, LIRIS UMR5205, F-69621 Villeurbanne, France
affil2 Univ Lyon, Université Claude Bernard Lyon 1, LIRIS UMR5205,
F-69622 Villeurbanne, France
http://liris.cnrs.fr/
prénom.nom@liris.cnrs.fr

**Résumé.** Dans un contexte de traitement de flux de données, il est important de garantir à l'utilisateur des propriétés de performance, qualité des résultats et passage à l'échelle. Mettre en adéquation ressources et besoins, pour n'allouer que les ressources nécessaires au traitement efficace des flux, est un défi d'actualité majeur au croisement des problématiques du Big Data et du Green IT. L'approche que nous suggérons permet d'adapter dynamiquement et automatiquement le degré de parallélisme des différents opérateurs composant une requête continue selon l'évolution du débit des flux traités. Nous proposons i) une métrique permettant d'estimer l'activité future des opérateurs selon l'évolution des flux en entrée, ii) l'approche *AUTOSCALE* évaluant a priori l'intérêt d'une modification du degré de parallélisme des opérateurs en prenant en compte l'impact sur le traitement des données dans sa globalité iii) grâce à une intégration de notre proposition à *Apache Storm*, nous exposons des tests de performance comparant notre approche par rapport à la solution native de cet outil.

## 1 Introduction

Avec la multiplication des sources de flux de données (capteurs, objets connectés...), les méthodes d'acquisition, stockage et traitement de ces données ont évolué pour en gérer la masse et la vélocité. Ces flux sont des séquences de n-uplets dont le débit et la distribution des valeurs peuvent varier au cours du temps. L'interrogation de ces flux via des requêtes, dites *continues* (Sattler et Beier (2013)), soulèvent des défis majeurs en terme de performance et passage à l'échelle. En terme de performance, les systèmes de gestion de flux de données doivent pouvoir traiter à la volée les données issues de flux. De la capacité de ces systèmes à absorber ces flux pour les traiter, dépend également la qualité des résultats qui seront produits. En ce qui concerne le passage à l'échelle, ces systèmes doivent être en mesure d'absorber des débits de données potentiellement très variables et élevés.

Travaux partiellement financés par le projet Socioplug ANR-13-INFR-0003, http://socioplug.univ-nantes.fr/index.php/SocioPlug_Project