# Délestage pour l'analyse multidimensionnelle de flux de données

Sylvain Ferrandiz, Georges Hébrail

GET / Télécom Paris
46, rue Barrault
F-75634 Paris Cedex 13
sylvain.ferrandiz@enst.fr
georges.hebrail@enst.fr

**Résumé.** Dans le contexte de la gestion de flux de données, les données entrent dans le système à leur rythme. Des mécanismes de délestage sont à mettre en place pour qu'un tel système puisse faire face aux situations où le débit des données dépasse ses capacités de traitement. Le lien entre réduction de la charge et dégradation de la qualité des résultats doit alors être quantifié.

Dans cet article, nous nous plaçons dans le cas où le système est un cube de données, dont la structure est connue a priori, alimenté par un flux de données. Nous proposons un mécanisme de délestage pour les situations de surcharge et quantifions la dégradation de la qualité des résultats dans les cellules du cube. Nous exploitons l'inégalité de Hoeffding pour obtenir une borne probabiliste sur l'écart entre la valeur attendue et la valeur estimée.

## 1 La gestion de flux de données

Les avancées de l'électronique et de l'informatique enrichissent continuellement la pratique de la récolte et de la gestion des données. La constante est l'accroissement des capacités de traitement, tant au niveau de l'acquisition que du stockage et de l'accès aux données. Mais lorsque l'information doit être extraite instantanément de données récoltées continuellement, le modèle relationnel basé sur des tables atteint ses limites. C'est là qu'interviennent les flux de données.

Un flux de données est une suite de tuples ayant tous la même structure. Cette structure est représentée par un schéma, comprenant le nom des champs du tuple et leur type. La différence entre un flux et une table est le caractère ordonné des tuples. L'ordre est souvent déterminé par un champ d'agencement (typiquement la date, mais pas nécessairement). On entre dans le cadre de la gestion de flux dès lors que

- les données du flux n'ont pas vocation à être stockées,
- les données nécessitent un traitement immédiat,
- les requêtes sont exécutées continuellement (*i.e.* les flux de données donnent naissance à d'autres flux de données).

La gestion de flux de données repose sur un modèle "data push" : les données se présentent d'elles-mêmes, à leur propre rythme. En conséquence, le système ne maîtrise pas et ne connaît