# Echantillonnage spatio-temporel de flux de données distribués

Raja Chiky*, Jérôme Cubillé**, Alain Dessertaine**,
Georges Hébrail*, Marie-Luce Picard **

* GET-ENST Paris
Laboratoire LTCI - UMR 5141 CNRS - Département Informatique et Réseaux
46 rue Barrault, 75634 Paris Cedex 13
Email: prenom.nom@enst.fr
** EDF R&D - Départements ICAME et OSIRIS
1, Avenue du Général de Gaulle, 92140 Clamart
Email: prenom.nom@edf.fr

**Résumé.** Ces dernières années, sont apparues de nombreuses applications, utilisant des données potentiellement infinies, provenant de façon continue de capteurs distribués. On retrouve ces capteurs dans des domaines aussi divers que la météorologie (établir des prévisions), le domaine militaire (surveiller des zones sensibles), l'analyse des consommations électriques (transmettre des alertes en cas de consommation anormale),... Pour faire face à la volumétrie et au taux d'arrivée des flux de données, des traitements sont effectués 'à la volée' sur les flux. En particulier, si le système n'est pas assez rapide pour traiter toutes les données d'un flux, il est possible de construire des résumés de l'information. Cette communication a pour objectif de faire un premier point sur nos travaux d'échantillonnage dans un environnement de flux de données fortement distribués. Notre approche est basée sur la théorie des sondages, l'analyse des données fonctionnelles et la gestion de flux de données. Cette approche sera illustrée par un cas réel : celui des mesures de consommations électriques.

## 1 Motivations

Les entrepôts de données sont de plus en plus alimentés par des flux de données provenant d'un grand nombre de capteurs distribués. Malgré l'évolution des nouvelles technologies de traitement et de stockage des données, il reste difficile voire impossible de conserver la totalité de l'information. Pour faire face à cette inflation, de nombreux travaux (Aggarwal, 2007; Babcock et al, 2002; Muthukrishnan, 2005) ont été menés ces dernières années sur la gestion et l'analyse de flux de données : un flux de données est défini comme une séquence continue, potentiellement infinie, de n-uplets (d'enregistrements) ayant tous la même structure. L'ordre d'arrivée des n-uplets n'est pas contrôlé, et les données, de par l'importance de leur volume et de leur débit d'arrivée, ne peuvent pas exhaustivement être stockées sur disque : les données passent, et doivent être traitées 'à la volée'.