Echantillonnage optimisé de données temporelles distribuées pour l'alimentation des entrepôts de données

Raja Chiky*, Georges Hébrail*,**

* 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 - Département ICAME
1, Avenue du Général de Gaulle, 92140 Clamart
Email: georges.hebrail@edf.fr

Résumé. Les entrepôts de données sont de plus en plus alimentés par des données provenant d'un grand nombre de capteurs. Les capteurs trouvent leur utilité dans plusieurs domaines : médical, militaire, trafic routier, météorologie ou encore des données de consommation électrique. 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 avant leur enregistrement dans les entrepôts de données. Nous présentons des algorithmes d'échantillonnage optimisé sur des flux de données provenant de capteurs distribués. L'efficacité des algorithmes proposés a été testée sur un jeu de données de consommation électrique.

1 Introduction

Les entrepôts de données (data warehouses) sont utilisés afin d'améliorer la prise de décision dans les entreprises. Ils servent à historiser des données résumées, non volatiles et disponibles pour l'interrogation, l'analyse et la prise de décision. Les entrepôts de données sont de plus en plus alimentés par des données provenant d'un grand nombre 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),... Le concepteur d'un entrepôt de données doit mettre en place une stratégie de mise à jour pour l'historisation en prenant en compte la volumétrie des données et en garantissant les meilleures performances possibles pour l'entrepôt en terme de temps de réponse pour l'interrogation et l'analyse.
Cet article traite des problèmes liés à des données temporelles et distribuées, mesurées en temps réel, où les sources des données correspondent à un grand nombre de capteurs qui enregistrent périodiquement des mesures dans des domaines spécifiques (température, consommation électrique...). La figure 1 montre un exemple d'architecture de récupération de données à partir de capteurs servant à mesurer des index de consommation électrique. Il s'agit de plusieurs millions de compteurs électriques communicants qui sont reliés à des concentrateurs, qui à leur tour sont reliés à des entrepôts de données. L'envoi des données se fait sous forme de flux,