Interrogation des résumés de flux de données

Nesrine Gabsi*,**, Fabrice Clérot **
Georges Hébrail*

*Institut TELECOM ; TELECOM ParisTech ; CNRS LTCI
46, rue Barrault 75013 Paris
prénom.nom@telecom-paristech.fr,
** France Telecom RD
2, avenue P.Marzin 22307 Lannion
prénom.nom@orange-ftgroup.com

Résumé. Les systèmes de gestion de flux de données (SGFD) ont été conçus afin de traiter une masse importante de données produites en ligne de façon continue. Etant donné que les ressources matérielles ne permettent pas de conserver toute cette volumétrie, seule la partie récente du flux est mémorisée dans la mémoire du SGFD. Ainsi, les requêtes évaluées par ces systèmes ne peuvent porter que sur les données les plus récentes du flux. Par conséquent, les SGFD actuels ne peuvent pas traiter des requêtes qui portent sur des périodes très longues. Nous proposons dans cet article, une approche permettant d'évaluer des requêtes qui portent sur une période plus longue que la mémoire du SGFD. Ces fenêtres font appels à des données récentes et des données historisées. Nous présentons le niveau logique de cette approche ainsi que son implantation sous le SGFD Esper. Une technique d'échantillonnage associée à une technique de fenêtre point de repère est appliquée pour conserver une représentation compacte des données du flux.

1 Introduction

Contrairement aux systèmes transactionnels classiques où les données sont conservées avant d'être traitées, les Systèmes de Gestion de Flux de Données (SGFD)(Babcock et al. (2002a)) effectuent un traitement à la volée en une seule passe (sans conservation a priori des données) et permettent de poser des requêtes continues. Compte tenu du volume et du débit des données, les SGFD ne conservent que les données récentes ou celles du passé récent dans des structures appelées *fenêtres*. Ces données ne peuvent être analysées a posteriori. Il est ainsi nécessaire, pour tout besoin d'analyse sur les flux, de préciser a priori sa tâche avant l'arrivée des données. Cependant, si un nouveau besoin portant sur des données écoulées est exprimé, ces tâches ne sont plus réalisables. Une solution permettant le traitement a posteriori consiste à conserver un résumé du flux. Il s'agit de résumer le contenu du flux de façon à construire un modèle résumé qui, bien que beaucoup plus petit en taille, permet de répondre à ces tâches mais d'une manière approchée.