Résumé généraliste de flux de données

Projet MIDAS ANR07-MDCO-008, financé par l'ANR
Collectif d'auteurs participant au projet MIDAS*

*contacts : Georges Hébrail ou Christine Potier
Télécom ParisTech, Département INFRES
46 rue Barrault, 75634 Paris Cedex 13
{georges.hebrail, christine.potier}@telecom-paristech.fr
http://midas.enst.fr/

Résumé. Lorsque le volume des données est trop important pour qu'elles soient stockées dans une base de données, ou lorsque leur fréquence de production est élevée, les Systèmes de Gestion de Flux de Données (SGFD) permettent de capturer des flux d'enregistrements structurés et de les interroger à la volée par des requêtes permanentes (exécutées de façon continue). Mais les SGFD ne conservent pas l'historique des flux qui est perdu à jamais. Cette communication propose une définition formelle de ce que devrait être un résumé généraliste de flux de données. La notion de résumé généraliste est liée à la capacité de répondre à des requêtes variées et de réaliser des tâches variées de fouille de données, en utilisant le résumé à la place du flux d'origine. Une revue de plusieurs approches de résumés est ensuite réalisée dans le cadre de cette définition.

1 Introduction

Les nombreux travaux récents relatifs aux flux de données ont permis de dégager clairement les grandes différences entre "traitement des flux de données" et "traitement des bases de données". Une des différences structurantes est que le traitement des flux de données repose sur l'exécution de requêtes continues (valables sur la durée du flux) sur des données volatiles (potentiellement infinies par nature, le flux ne peut pas être stocké) alors que le traitement des bases de données repose sur l'exécution de requêtes volatiles ("one shot") sur des données persistantes (Golab et Özsu 2003).

A l'évidence, une requête continue sur un flux de données ne peut fournir de réponse que sur une période nécessairement limitée. Il est dans la nature des données volatiles de n'être plus disponibles pour les analyses après leur expiration, ce qui suppose de poser a priori sur un flux l'ensemble des requêtes dont on pourra avoir besoin par la suite. Il est clairement hors de question de poser des requêtes a posteriori sur le flux. Ici "a posteriori" doit être entendu comme portant sur des données ayant expiré.

Pour pallier cet inconvénient, différents algorithmes et structures de données ont été proposés dans la littérature, souvent (mais pas toujours) sous le nom de "résumé" ("summary") sans toutefois qu'émerge une notion cohérente de ce qu'est ou devrait être le résumé d'un flux de données.

Dans cet article, nous donnons d'abord la définition de "résumé généraliste" et nous analysons quelques techniques de résumé de flux en regard de cette définition. Puis dans le paragraphe 3, nous analysons le traitement de la dimension temporelle dans chacune de ces méthodes.