

Fragmentation des entrepôts de données XML

Hadj Mahboubi et Jérôme Darmont

Université de Lyon (ERIC Lyon 2)
5 avenue Pierre Mendès-France
69676 Bron Cedex
{hmahboubi,jdarmont}@eric.univ-lyon2.fr

Résumé. Les entrepôts de données XML proposent une base intéressante pour les applications décisionnelles qui exploitent des données hétérogènes et provenant de sources multiples. Cependant, les performances des SGBD natifs XML étant actuellement limitées en termes de temps de réponse et de volume des données, il est nécessaire de trouver des moyens pour les optimiser. Dans cet article, nous proposons une adaptation de la fragmentation horizontale dérivée définie dans le contexte relationnel aux entrepôts de données XML pour traiter ces deux problèmes. Nos expérimentations montrent que cette démarche permet de réduire le temps de traitement de requêtes décisionnelles XQuery de façon significative.

Mots clés : Entrepôts de données XML, performance, fragmentation.

1 Introduction

Actuellement, les applications décisionnelles exploitent de plus en plus de données hétérogènes et provenant de sources variées. Dans ce contexte, le langage XML peut grandement aider à l'intégration et à l'entreposage de données en vue de la fouille ou de l'analyse statistique ou en ligne (Beyer et al., 2005). Cependant, les requêtes décisionnelles sont généralement complexes du fait qu'elles impliquent de nombreuses jointures et agrégations. Par ailleurs, les systèmes de gestion de bases de données (SGBD) natifs XML présentent actuellement des performances médiocres quand les volumes de données sont importants ou que les requêtes sont complexes. Il est donc crucial lors de la construction d'un entrepôt de données XML de garantir la performance des requêtes XQuery qui l'exploiteront.

La fragmentation est une technique d'optimisation de performance qui consiste à diviser un ensemble de données en plusieurs fragments de telle façon que la combinaison de ces fragments produit l'intégralité des données source, sans perte ou ajout d'information. Dans le contexte relationnel, ce sont les tables (relations) qui sont partitionnées ; dans le contexte XML, ce sont les documents XML. La fragmentation peut intégrer plusieurs scénarios : la conception de systèmes distribués, le traitement des flux de données et l'échange de données dans les systèmes pair-à-pair.

Les travaux qui traitent de la fragmentation dans les entrepôts de données relationnels s'inspirent de ceux proposés dans les bases de données relationnelles et orientées objets. Cette