Une Démarche Conjointe de Fragmentation et de Placement dans le Cadre des Entrepôts de Données Parallèles

Soumia Benkrid *, Ladjel Bellatreche **

* Ecole Nationale Supérieure en Informatique, Oued-Smar Alger - Algérie
s_benkrid@esi.dz
** LISI/ENSMA - Université de Poiters, Futuroscope 86960 France
bellatreche@ensma.fr

Résumé. Traditionnellement, concevoir un entrepôt de données parallèle consiste d'abord à partitionner son schéma ensuite allouer les fragments générés sur les noeuds d'une machine parallèle. L'inconvénient majeur d'une telle approche est son ignorance de l'interdépendance entre les processus de fragmentation et d'allocation. Une des entrées du problème d'allocation est l'ensemble de fragments générés par la fragmentation. Notons que les deux processus cherchent à optimiser le même ensemble de requêtes. Dans ce papier, nous proposons une approche de conception d'un entrepôt de données relationnel parallèle selon une architecture distribuée (shared nothing) intégrant les processus de fragmentation et d'allocation. Ensuite, une méthode de répartition de charges sur les noeuds de la machine parallèle est proposée. Finalement, une validation de nos propositions en utilisant le banc d'essai APB-1 release II est présentée.

1 Introduction

Le processus d'aide à la décision est souvent mené par l'intermédiaire de requêtes complexes caractérisées par des jointures, sélections et agrégations exécutées sur des entrepôts de données très volumineux. Au cours de la dernière décennie, la taille des entrepôts de données a augmenté de 5 à 100 téra octets (DeWitt et al.). Pour optimiser les requêtes décisionnelles sur des entrepôts de données de telle taille, les techniques d'optimisation classiques telles que les *vues matérialisées*, les *index avancés*, la *fragmentation* ne sont pas suffisantes. Le traitement parallèle devient alors une solution incontournable pour réduire les coûts de requêtes complexes. Ce dernier n'a pas eu la même attention de la part de la communauté des entrepôts de données contrairement aux techniques classiques ; à l'exception des travaux de (Stöhr et al., 2000; Stöhr et Rahm, 2001) et de (Furtado, 2004). Paradoxalement, les éditeurs de bases de données proposent des solutions parallèles comme InfoSphere d'IBM et Oracle.

La conception d'un entrepôt de données parallèle passe par cinq phases principales : (i) le choix de l'architecture matérielle, (ii) la fragmentation de l'entrepôt de données, (iii) l'allocation (ou le placement) de fragments générés par le processus de la fragmentation, (vi) la répartition des charges et (v) le traitement des requêtes.