Entrepôts de données sur grilles de calcul

Pascal Wehrle, Maryvonne Miquel, Anne Tchounikine
LIRIS UMR 5205
INSA, Campus de la Doua,
Bâtiment Blaise Pascal (501), 20, avenue Albert Einstein
69621 VILLEURBANNE CEDEX
<prenom>.<nom>@insa-lyon.fr
http://liris.cnrs.fr

1 Introduction

L'objectif d'un entrepôt de données est de mettre à disposition des outils d'aide à la décision à partir de grands volumes de données produits par des systèmes d'informations de production (Inmon 1992). Les « dimensions » du modèle multidimensionnel représentent les axes d'analyse et sont hiérarchisées en niveaux de détail. Les données sont modélisées sous forme d'hypercubes navigables grâce aux outils OLAP (On Line Analytical Processing). La structure interne classique d'un entrepôt de données est celle du schéma en étoile, introduit par Kimball (Kimball 1996). Celui-ci est constitué d'une table de faits centrale contenant les données les plus détaillées de l'entrepôt, appelées « faits » ou « mesures ». Celles-ci sont associées via des clés étrangères à des tables de dimension accueillant les données concernant les axes d'analyse. Afin d'améliorer les temps de réponse aux requêtes, des agrégats comme par exemple la somme ou la moyenne sur les faits sont pré-calculés au sein de l'entrepôt.

Les besoins croissants en termes de capacité de traitement et de stockage causés par la conception et l'exploitation d'entrepôts de données de plus en plus complexes et volumineux par exemple dans le secteur geno-médical (Brunie et al. 2003) favorisent l'utilisation de systèmes distribués puissants. Le concept récent des grilles de calcul fournit une approche décentralisée à la construction d'infrastructures à hautes performances efficientes, économiques et extensibles dont les principes de base sont exposés par Foster (Foster 2003). Leurs services de gestion et d'information mettent à disposition un accès transparent à un grand nombre de ressources hétérogènes distantes dans le but d'offrir à l'utilisateur une qualité de service « non triviale ».

Le modèle d'architecture proposé dans cet article a pour objectif l'intégration d'un entrepôt de données sur une infrastructure de grille de calcul. Les avantages principaux d'un déploiement de grands volumes de données détaillées et de leurs agrégats sur une grille sont les possibilités de traitement et d'accès parallèles, de stockage et d'échange décentralisés des données ou résultats de requêtes. Du coté utilisateur, l'objectif est de proposer un service d'entrepôt aux spécialistes connectés à différents points d'accès de la grille.

2 Identification et fragmentation des données

L'entrepôt de données doit être entièrement réparti parmi les noeuds de la grille afin de s'adapter à l'infrastructure de grille de calcul et de permettre une gestion et un accès décentralisé. Pour faciliter la recherche et l'échange de données entre noeuds de la grille nous introduisons une méthode d'identification unique et globale des données de l'entrepôt.
Des identifiants uniques pour les données les plus détaillées sont facilement trouvés grâce aux membres de dimension directement associés aux faits. Il est important pour une gestion efficace des ensembles de données de pouvoir ordonner et comparer ces identifiants.