Entrepôts de données multidimensionnelles réduites : principes et expérimentations

Faten Atigui, Franck Ravat, Jiefu Song, Gilles Zurfluh

IRIT (UMR 5505) Institut de Recherche en Informatique de Toulouse
118 route de Narbonne, F-31062 Toulouse, France
{atigui, ravat, song, zurfluh}@irit.fr

Résumé. Notre objectif est de proposer une solution pour la réduction de données d'un Entrepôt de Données Multidimensionnelles (EDM) afin d'obtenir des schémas agrégés sur différentes périodes et de ne retenir que les informations pertinentes pour les prises de décision. Dans un premier temps, nous proposons une solution pour la modélisation des EDM réduits basée sur des états contenant des schémas en étoile et des opérateurs de réduction pour définir les schémas réduits. Dans un second temps, nous décrivons nos expérimentations et les résultats obtenus dans différents contextes : BD R-OLAP sans réduction et BD R-OLAP réduite. Nous montrons que, quelques soit le type d'analyse, les exécutions dans un contexte réduit sont plus performantes.

1 Introduction

Les systèmes d'aide à la décision sont généralement supportés par des Entrepôts de Données Multidimensionnelles (EDM). Un schéma d'EDM est basé sur des faits (sujets d'analyse) et des dimensions (axes d'analyses). Les faits contiennent des indicateurs tandis que les dimensions contiennent les paramètres d'analyse. Ces paramètres sont organisés en hiérarchies permettant de classer les paramètres du niveau de granularité minimale vers le niveau de granularité maximale. Par définition, un EDM stocke de manière permanente les données décisionnelles et ces données sont régulièrement mises à jour (ajouts de nouvelles valeurs). De ce fait, un EDM gère des volumes de données de plus en plus importants, ce qui entraîne de nombreux problèmes en termes de performance et de stockage. Cependant, la pertinence d'une donnée décisionnelle est susceptible de décroître avec l'âge : les informations détaillées sont essentielles pour des données les plus récentes mais une information plus agrégée est souvent suffisante pour des données plus anciennes Skyt et al. (2008). Par exemple, un responsable marketing analyse les ventes à un niveau de granularité permettant de manipuler chaque produit durant les 2 ou 3 années les plus récentes mais ce niveau de granularité peut s'avérer inutile pour une période plus ancienne (la plupart de ces produits n'existant plus) ; le décideur pourra analyser ces ventes non plus par produit, mais au niveau plus général de la catégorie des produits qui reste stable dans le temps. Faisant face à ce grand nombre de données dont une grande partie n'est pas pertinente pour la prise de décision, notre objectif est à la fois d'augmenter la performance de traitement des requêtes portant sur un gros volume de données et