Archivage d'entrepôts de données multidimensionnelles

Faten Atigui, Franck Ravat, Olivier Teste, Gilles Zurfluh

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

Résumé. Les données d'un entrepôt sont rafraîchies périodiquement et conservées de manière permanente. Cependant, les décideurs portent généralement un intérêt moindre pour les données anciennes. Dans cet article, nous proposons un mécanisme permettant de synthétiser les données les plus anciennes. Nous définissons un modèle conceptuel d'archivage de données multidimensionnelles. Nous présentons, ensuite, le modèle logique correspondant et les principes permettant d'interroger des schémas multidimensionnels archivés.

1 Introduction

Un entrepôt de données (ED) est une collection de données thématiques, intégrées, non volatiles et historisées pour des fins décisionnelles. Les données pertinentes pour la prise de décision sont collectées à partir des sources par le biais des processus d'Extraction-Transformation-Chargement. Dans un ED, les données extraites sont souvent structurées selon un format multidimensionnel (MD) qui organise l'information en termes de sujets d'analyse (faits) et d'axes d'analyse (dimensions) au sein d'un schéma en étoile Kimball (1996).

Dans un ED, les données sont conservées de manière permanente et sont rafraîchies de manière récurrente. De ce fait, l'ED possède de gros volume de données dans lequel le décideur risque de « se perdre » lors de ses analyses. De plus, les données historisées perdent de leur intérêt avec le temps : alors que la granularité des informations doit généralement être importante pour des données récentes ; elle peut être plus faible pour des données anciennes. Par exemple, un décideur peut analyser ses ventes au niveau de la granularité produit sur les cinq dernières années tandis que pour les périodes antérieures, ces analyses au niveau du produit seraient absurdes (les produits n'existent plus à l'heure actuelle) et donc le décideur fera des analyses au niveau de la gamme du produit (qui n'a pas évolué dans le temps). Afin de faciliter la tâche du décideur et de mieux répondre à ses besoins, il est préférable de garder uniquement l'information nécessaire à ses analyses. L'idée est donc d'offrir un environnement d'analyse MD adapté aux besoins des décideurs en leur permettant de supprimer dans le temps les niveaux de granularité inutiles pour leurs analyses.

Notre objectif est donc de proposer un modèle de données MD permettant de représenter l'archivage de données afin de ne conserver que les données nécessaires aux analyses décisionnelles. Ce mécanisme permet de garder une vue synthétique sur les données les moins récentes et dont les détails sont superflus pour la prise de décision. De plus, cette solution permettrait d'anticiper le problème de temps de réponse aux requêtes MD dès les premières phases