

Modèle unifié pour la transformation des schémas en constellation

Faten Atigui^{*,**}, Franck Ravat^{*,**}, Olivier Teste^{*,***}, Gilles Zurfluh^{*,**}

^{*}IRIT (UMR 5505), Institut de Recherche en Informatique de Toulouse,
118 route de Narbonne, 31062 Toulouse, France

^{**}Université Toulouse 1 Capitole, 2 rue G. Marty, 31042 Toulouse Cedex9, France

^{***}Université Toulouse 3 Paul Sabatier 118 route de Narbonne, 31062 Toulouse, France
{atigui, ravat, teste, zurfluh}@irit.fr

Résumé. Au cours de ces dernières années, plusieurs approches ont abordé la modélisation et le développement des Entrepôts de Données (ED). La plupart de ces approches fournit des solutions partielles qui traitent soit la modélisation multidimensionnelles, soit la modélisation des processus d'Extraction-Transformation-Loading (ETL). Toutefois, peu de travaux ont visé à unifier ces deux problématiques dans un cadre structuré ou à automatiser le processus d'entrepotage. Afin de pallier ces limites, nous proposons dans ce papier une démarche unifiée et semi-automatique qui intègre la modélisation des ED et des processus ETL. Cette démarche est définie dans le cadre d'une Architecture Dirigée par les Modèles (MDA). Elle permet (i) de formaliser les besoins des décideurs, ensuite (ii) de générer les modèles conceptuel, logique et physique de l'ED et des processus ETL conjoints (iii) ainsi que les codes de création et d'alimentation (ETL) des structures multidimensionnelles. Les règles de transformations entre modèles sont formalisées en Query/View/Transformation.

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 au moyen des processus d'Extraction-Transformation-Chargement couramment connus sous le nom ETL (Extraction-Transformation-Loading) (Vassiliadis, 2009). Les données extraites sont souvent structurées selon un format multidimensionnel qui organise l'information en termes de faits et de dimensions (Kimball, 1996).

De manière générale, le processus d'entrepotage se fait en deux temps. Tout d'abord, il faut créer les structures multidimensionnelles de l'ED. Ensuite, il faut extraire les données des sources pour alimenter l'ED. La première phase repose sur trois étapes principales, à savoir la modélisation conceptuelle, la modélisation logique et la modélisation physique. La modélisation conceptuelle vise à fournir une représentation commune des données provenant de diverses sources opérationnelles. Le schéma conceptuel est transformé en un schéma logique puis en un schéma physique spécifique à une plateforme donnée (Romero et Abelló, 2009). La seconde phase vise à décrire et à implanter les processus ETL. Pour ce faire, il est