ProtOLAP : un système de prototypage rapide pour les entrepôts de données

Hassan Nazih**, Myoung-Ah Kang*, Sandro Bimonte**

*LIMOS-CNRS, ISIMA, Blaise Pascal University, Campus des Cezeaux Aubière, France
kang@isima.fr
**IRSTEA, TSCF 24 Av. Des Landais Aubière, France
hassanazih@gmail.com, sandro.bimonte@irstea.fr

Résumé. Les approches disponibles pour concevoir un entrepôt de données, y compris celles adoptant les pratiques agiles, sont basées sur l'hypothèse que les données sources sont connues et disponibles à l'avance. Cette hypothèse n'est pas toujours vraie dans certains contextes. Pour pallier ces limites, nous proposons ProtOLAP, une méthodologie assistée par un outil de prototypage rapide.

1 Introduction

Les Entrepôts de données (EDs) visent à soutenir le processus décisionnel en permettant des analyses flexibles et interactives (Kimball, 2008). Les clients OLAP permettent aux décideurs de visualiser et d'explorer des données en appliquant les opérateurs OLAP. Une distinction peut être faite sur les méthodologies de conception de ED selon le rôle des besoins de utilisateurs ; les approches axées sur les besoins ("requirement-driven approaches"), les approches axées sur les sources ("source-driven approaches") et les approches mixtes. Toutes ces approches sont basées sur l'hypothèse que les données sources pour alimenter l'ED sont connues et disponibles à l'avance. Toutefois, dans certains cas les données sources ne sont pas disponibles au début du projet. C'est le cas par exemple de la partie 'Irrigation' du projet national français EDEN pour l'analyse des consommations d'électricité et d'eau dans les fermes agricoles (Boulil et al., 2015).

Dans ce projet, les données sources sont identifiées et recueillies a posteriori, selon les besoins d'information représentés dans le schéma conceptuel. Les décideurs impliqués dans ce projet non expert en TIC (Technologies de l'Information et de la Communication) ont jugé aussi très difficile d'exprimer leurs besoins d'analyse par des modèles conceptuel. Ceci a rendu le processus de validation des besoins utilisateur plus long et incertain. Nous avons résolu ce problème en leur montrant des exemples de résultats de requêtes OLAP possibles au lieu de travailler sur les schémas pour identifier leurs besoins d'analyses. Nous proposons ainsi une méthodologie assistée par un outil nommée ProtOLAP (Bimonte et al., 2013) qui permet des tests rapides et fiables pour la validation des schémas de EDs dans des situations où les compétences en TIC des utilisateurs sont minimes et les données sources ne sont pas disponible au début de projet.

2 Etat de l'art

Des modèles conceptuels (Romero et Abellò, 2009) et modèles de spécification des besoins pour EDs (Souza et al., 2012) ont été proposés dans la littérature pour aider les concepteurs et les décideurs à discuter et à valider les exigences au cours des développements. Dans Huynh et Schiefer (2001), les auteurs présentent un outil pour alimenter un ED avec les exemples de données obtenues à l'aide de méthodes statistiques appliquées aux données sources. Toutefois, cet outil n'est pas intégré au sein d'un système complet pour le prototypage rapide, donc il n'est pas directement utilisable dans notre étude de cas. Enfin, l'infrastructure de test de l'ED Golfarelli et Rizzi (2011) propose un test avec des échantillons de données sources durant les premières phases de conception. Ceci vise à permettre aux utilisateurs de valider les exigences lors de la conception, afin de réduire le coût de correction des erreurs et des "malentendus". Cependant il nécessite que certaines données sources soient disponibles, ce qui n'est pas notre cas.

3 ProtOLAP

Dans cette section, nous décrivons le système ProtOLAP qui implémente notre méthodologie par les étapes suivantes :

1. Les décideurs discutent de leurs besoins d'analyses avec les concepteurs, principalement en ce qui concerne les indicateurs dont ils ont besoin. Ensuite, les concepteurs peuvent élaborer rapidement un schéma conceptuel en utilisant le profil UML pour les entrepôts de données spatiales.
2. Lorsqu'une première version de schéma conceptuel est produite, la phase suivante génère un prototype. Dans cette étape, un schéma logique (relationnel) et les métadonnées associées sont automatiquement générés et déployés à partir du schéma conceptuel. La cible de SGBDR est Oracle, alors que la cible serveur OLAP est Mondrian.
3. Ensuite les décideurs peuvent insérer des données (réalistes) dans le prototype au moyen d'une interface adaptée.
4. Enfin, les décideurs peuvent explorer les données qu'ils ont insérées avec le client OLAP. En visualisant les exemples des requêtes, ils peuvent ainsi vérifier si le prototype (et par conséquent le schéma conceptuel sous-jacent) correspond à leurs besoins d'analyses. Si le prototype ne répond pas aux besoins des décideurs alors, la méthodologie recommence à lŠétape 1, autrement lŠimplémentation de lŠETL peut être faite.

L'architecture de ProtOLAP repose sur une plate-forme ROLAP et se compose de quatre niveaux : i) le *niveau des besoins* est utilisé pour produire des schémas conceptuels UML au moyen de l'outil MagicDraw ; ii) le *niveau de déploiement* comprend le SGBDR Oracle, le serveur OLAP Mondrian et l'outil développé pour créer les schémas relationnels pour Oracle et les métadonnées pour Mondrian à partir du schéma conceptuel produit par le niveau des besoins ; iii) le *niveau d'alimentation* génère automatiquement une interface visuelle grâce auquel les décideurs peuvent alimenter l'ED produit par le niveau de déploiement ; iv) le *niveau d'analyse* permet aux décideurs d'interroger les données stockées dans l'ED avec le client OLAP JRubik. Ces niveaux sont décrits en détail dans les sous-sections suivantes.

La méthode ProtOLAP est itérative, c'est-à-dire qu'elle permet d'affiner progressivement un schéma conceptuel et le prototype associé. Ainsi, le niveau des besoins doit aider efficacement les concepteurs à définir facilement et rapidement des schémas conceptuels multidimensionnels qui peuvent être déployés automatiquement. Pour ces raisons, nous avons choisi d'adopter le profil Spatial Datacube UML Boulil et al. (2015) et son implémentation sous MagicDraw. Ce profil étend des stéréotypes UML pour les applications OLAP complexes, et il garantit au moyen d'un ensemble de contraintes OCL que les schémas multidimensionnelles soient bien formés (par exemple une dimension doit avoir au moins un niveau).

Dans le niveau de déploiement, les données sont stockées sous Oracle Spatial. Mondrian est utilisé comme un serveur OLAP. Mondrian est un serveur OLAP open source appartenant à la Suite de Intelligent Pentaho Business. Il utilise des métadonnées XML pour mapper un schéma multidimensionnel à un schéma relationnel hébergé sur n'importe quel SGBDR. ProtOLAP prend comme entrée le fichier XMI (XML Metadata Interchange) du diagramme de classes UML du schéma conceptuel. Ensuite, l'outil génère automatiquement i) les schémas relationnels en étoile ou en flocon décrits en SQL pour Oracle Spatial, ii) les fichiers XML pour les métadonnées de Mondrian représentant les élements spatio-multidimensionnels (faits, niveaux, etc.) et les membres calculés en MDX pour les indicateurs complexes, iii) une représentation visuelle des éléments (hiérarchies) de l'ED générée sous forme d'arbre. Cette phase est complètement automatisée sans aucune intervention de l'utilisateur.

Le niveau d'alimentation offre un prototype avec une interface visuelle pour alimenter les données (les membres de dimensions et les mesures de faits) dans l'ED généré au niveau de déploiement. Les membres de dimensions peuvent être insérés manuellement ou automatiquement. Étant donné que les membres des dimensions sont organisés en hiérarchies, la vérification des relations hiérarchiques entre les membres est faite automatiquement. Les décideurs peuvent insérer des données sans aucune connaissance du modèle multidimensionnel. Les niveaux d'agrégation et des relations hiérarchiques sont intuitivement représentés par l'interface utilisateur. Une fois que toutes les dimensions sont alimentées, la table de faits est remplie avec des valeurs de mesures générées pour toutes les combinaisons possibles de membres de dimensions.

En fin, le niveau d'analyse est implémentée à l'aide de JRubik, un client Java développé sous Mondrian et compatible avec n'importe quel SGBDR, comme Oracle. En utilisant JRubik, les décideurs peuvent accéder et explorer d'une façon interactive et en quelques clicks, les données stockées au niveau d'alimentation. Des tableaux croisés permettent aussi aux décideurs de vérifier facilement si le modèle d'ED correspond à leur besoins, en affichant des données agrégées et en changeant les niveaux des dimensions dans le prototype. Par exemple, ils peuvent afficher la consommation totale d'eau par parcelles et mois en agrégeant la mesure correspondante avec la fonction sum sur les dimensions temporelle et spatiale, afin de savoir si ce genre de données d'agrégation correspond à leurs besoins.

Une vidéo démontrant le processus d'utilisation de ce système est mise en ligne sur le site, http ://www.isima.fr/˜kang/eda2014.html.

4 Conclusion

Dans cet article, nous avons proposé ProtOLAP, un outil et une méthode de prototypage rapide pour les projets d'EDs où les données sources ne sont pas disponibles au début et les

utilisateurs (décideurs) ont peu ou aucune compétence en TIC. ProtOLAP génère automatiquement un prototype simple qui permet d'insérer des données de test sans comprendre le modèle conceptuel ou multidimensionnel. Ceci aide les utilisateurs à voir si leurs besoins sont bien pris en compte dans le modèle conceptuel, en visualisant les exemples de requêtes OLAP avec les données du prototype. Nous travaillons pour proposer plusieurs versions de prototypes pour aider les concepteurs à suivre et gérer l'évolution du projet au cours de ses différentes itérations.

Références

Bimonte, S., E. Edoh-Alove, H. Nazih, M. Kang, et S. Rizzi (2013). Protolap : Rapid olap prototyping with on-demand data supply. In *International workshop on Data warehousing and OLAP, 28/10/2013-28/10/2013, San Francisco, USA*, pp. 61–66.

Boulil, K., S. Bimonte, et F. Pinet (2015). Spatial olap integrity constraints : from uml-based specification to automatic implementation : Application to energetic data in agriculture. *Journal of Decision Systems*.

Golfarelli, M. et S. Rizzi (2011). Data warehouse testing : A prototype-based methodology. *Information and Software Technology 53*, 1183–1198.

Huynh, N. et J. Schiefer (2001). Prototyping data warehouse systems. *In Proc. DaWaK*, 195–207.

Kimball, R. (2008). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. John Wiley and Sons.

Romero, O. et A. Abellò (2009). A survey of multidimensional modeling methodologies. *International Journal of Data Warehousing and Mining*, 1–23.

Souza, V., J. Mazòn, I. Garrigòs, J.Trujillo, et J. Mylopoulos (2012). Monitoring strategic goals in data warehouses with awareness requirements. *In Proc. Symposium on Applied Computing 2012*, 1075–1082.

Summary

The available approaches to data warehouse design, including those adopting agile practices, are based on the assumption that source data are known in advance and available. This hypothesis is not always true in some contexts. In this paper, we propose ProtOLAP, a tool-assisted fast prototyping methodology.