Entreposage de données complexes pour la médecine d'anticipation personnalisée

Jérôme Darmont

Université de Lyon (ERIC Lyon 2)
5 avenue Pierre Mendès-France
69676 Bron Cedex
France

jerome.darmont@univ-lyon2.fr

RÉSUMÉ. *L'utilisation généralisée des nouvelles technologies permet aujourd'hui d'envisager un suivi de santé tout au long de la vie, des traitements présymptomatiques, ainsi que des analyses variées sur des populations de patients données. Cette médecine basée sur l'information nécessite l'analyse de données diverses et hétérogènes : dossiers patients, images médicales, résultats d'analyses biologiques, etc. L'entreposage de données permet d'intégrer de telles données complexes et de les préparer pour des analyses en ligne, statistiques et/ou de fouille de données. Dans cet article, nous présentons un modèle d'entrepôt de données complexes relatives à des athlètes de haut niveau, dont l'objectif est de venir en support de la médecine d'anticipation personnalisée. Cet entrepôt est organisé comme une collection de magasins de données interconnectés par des dimensions communes.*

MOTS-CLÉS : *Entrepôts de données, magasins de données, données complexes, données biomédicales, médecine d'anticipation personnalisée.*

ABSTRACT. *With the growing use of new technologies, lifetime healthcare and pre-symptomatic treatment, as well as various analyses over a given population of patients, are becoming possible. Such information-based medicine requires the analysis of various and heterogeneous data, such as patient records, medical images, biological analysis results, etc. In this context, data warehousing represents an interesting solution for integrating what we term complex data and preparing them for statistical, on-line and/or data mining analyses. In this paper, we present the design of a complex data warehouse relating to high-level athletes that is aimed at supporting personalized, anticipative medicine. Our data warehouse is organized as a collection of interconnected datamarts sharing common dimensions.*

KEY WORDS: *Data warehousing, datamarts, complex data, biomedical data, personalized and anticipative medicine.*

1. Introduction

Avec l'usage généralisé des nouvelles technologies, le domaine de la santé connaît des évolutions importantes qui laissent imaginer la possibilité d'un suivi personnalisé des patients tout au long de leur vie, de traitements présymptomatiques ou encore d'analyses diverses concernant des populations de patients. Cette médecine basée sur l'information exploite des systèmes d'aide à la décision qui requièrent l'analyse de données variées et hétérogènes, comme des dossiers patients, des images médicales, des résultats d'analyses biologiques, etc. (Saad, 2004)

Les entrepôts de données (Inmon, 2002 ; Kimball *et al.*, 2002) peuvent former la base de tels systèmes décisionnels. Bien qu'ils soient conçus pour permettre l'analyse de données numériques, les concepts de l'entreposage de données demeurent valides pour ce que nous appelons les *données complexes*. Dans ce contexte, les mesures observées, bien que non nécessairement numériques, demeurent les indicateurs d'analyse. De plus, cette dernière s'effectue toujours selon différentes perspectives représentées par les dimensions de l'entrepôt. Les énormes volumes de données à traiter et leur historisation sont également des arguments en faveur d'une approche d'entreposage. Les entrepôts de données peuvent finalement être le socle de différents types d'analyses : statistiques, en ligne (OLAP – *On-Line Analytical Processing*) ou fouille de données.

Dans cet article, nous présentons un modèle d'entrepôt de données complexes relatives à la santé d'athlètes de haut niveau. L'objectif global de ce système est de faire des patients les gestionnaires de leur propre capital santé grâce à des recommandations d'entraînement, de nutrition, etc. L'originalité de cet outil est d'une part qu'il stocke des données médicales *complexes* issues de différents champs de la médecine et de la biologie, et d'autre part qu'il est conçu pour permettre deux types d'analyses innovants et sensiblement différents l'un de l'autre, afin de venir en support de : 1) la médecine d'anticipation personnalisée (MAP, en opposition à la médecine curative) pour des patients bien identifiés ; 2) des analyses statistiques à large spectre sur des populations de patients. Notre entrepôt est également conçu pour être évolutif et prendre en compte les futures avancées de la médecine.

Cet article est organisé comme suit. La Section 2 présente l'architecture globale de notre entrepôt de données. La Section 3 détaille le traitement de données complexes cardiovasculaires dans ce contexte. Nous concluons cet article et évoquons des perspectives de recherche dans la Section 4.

2. Architecture globale de l'entrepôt

Afin de rendre notre solution évolutive, nous avons adopté une architecture en bus décisionnel (Firestone, 2002 ; Kimball *et al.*, 2002). Notre entrepôt de données est composé d'un ensemble de dimensions conformes et de définitions de faits

standard. Dans ce contexte, les données relatives aux différentes spécialités médicales que nous devons prendre en compte représentent des magasins de données qui s'intègrent dans le bus décisionnel et bénéficient des dimensions et des tables de faits qui leurs sont nécessaires. L'union de ces magasins de données peut être vue comme l'entrepôt de données entier.

La Figure 1 présente l'architecture globale de notre entrepôt. Les carrés droits symbolisent les tables de faits, les carrés aux coins arrondis les dimensions, les pointillés entourent les différents magasins de données et le bus décisionnel est compris dans un rectangle arrondi grisé. Il est constitué par les dimensions communes à plusieurs magasins. Parmi elles, les dimensions principales communes à tous les magasins sont les patients, les fournisseurs de données, le temps et l'analyse médicale (qui regroupe plusieurs types d'analyse). Bien sûr, certains magasins de données (comme le magasin cardiovasculaire) possèdent des dimensions non partagées. Notre entrepôt inclut également un magasin des antécédents médicaux des patients qui n'est pas représenté sur la Figure 1, et d'autres magasins étaient prévus avant l'arrêt du projet MAP.

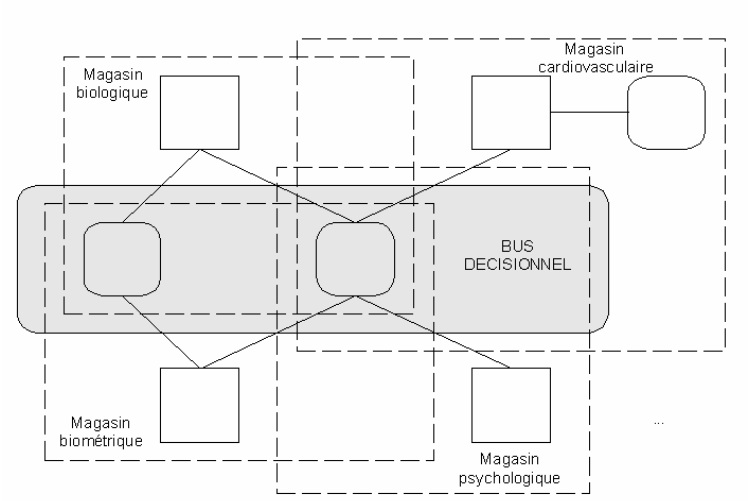

Figure 1. *Architecture globale de l'entrepôt de données MAP*

3. Magasin de données cardiovasculaires complexes

La Figure 2 représente l'architecture de notre magasin de données cardiovasculaires. La nature complexe des données sources, qui sont constituées de mesures brutes (taille des ventricules, par exemple), de documents multimédias (échocardiogrammes, par exemple) et de la conclusion du médecin quant à l'état de

santé du patient, ne peuvent pas être stockés dans une table de faits unique comme il est d'usage. C'est pourquoi nous exploitons un ensemble de tables interconnectées qui, ensemble, représentent les faits. Elles figurent sous la forme de carrés en pointillés dans la Figure 2.

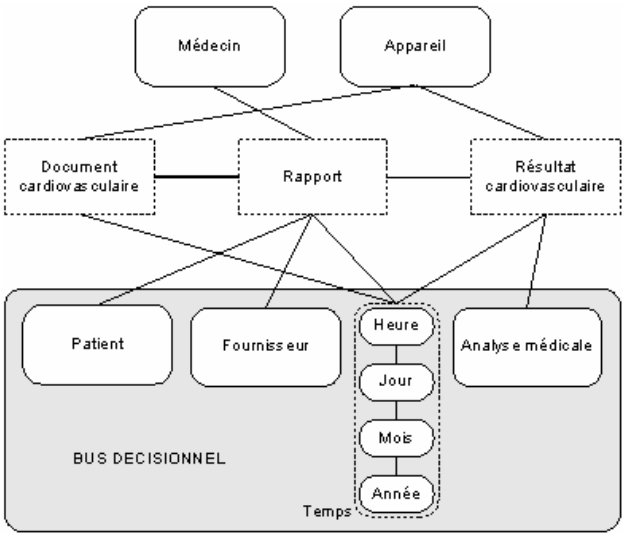

Figure 2. *Architecture du magasin de données cardiovasculaires*

Le rapport contient principalement la conclusion du médecin. C'est l'élément central dans notre « fait complexe ». Il est relié à plusieurs résultats d'analyse qui permettent de au médecin de tirer sa conclusion. Il est également relatif à des documents multimédias tels que des images médicales, qui aident aussi au diagnostic. Remarquons que cette association, représentée en gras dans la Figure 2, est une association « plusieurs-à-plusieurs ». Certains documents peuvent en effet être référencés par plusieurs rapports, par exemple pour prendre en compte l'évolution de l'état d'un patient au cours du temps grâce à une série d'échocardiogrammes. Chaque composant de notre « fait complexe » peut être individuellement relié aux dimensions. Les documents cardiovasculaires et les résultats ne sont en effet pas des descripteurs d'un éventuel fait « rapport », ils font partie d'un fait plus flou composé de plusieurs entités.

Finalement, remarquons que les documents cardiovasculaires ne peuvent pas actuellement être exploités lors d'analyses OLAP. Cependant, nous devons les stocker pour des raisons médicolégales. De plus, des extensions des opérateurs OLAP actuels permettant d'agréger des données non numériques (Ben Messaoud *et al.*, 2004) devraient rendre cette utilisation possible dans un proche futur.

4. Conclusion

Nous avons présenté dans cet article un entrepôt de données complexes venant en support de la médecine d'anticipation personnalisée. Son objectif est de permettre à la fois des analyses orientées patients et des études statistiques à large spectre. En adoptant une architecture en bus décisionnel, nous avons conçu notre entrepôt de manière globale afin de prendre en compte plusieurs spécialités médicales, et évolutive afin d'anticiper les avancées de la recherche médicale.

Les perspectives directes de ce travail sont doubles. La première concernait avant l'arrêt du projet le contenu de l'entrepôt, c'est-à-dire l'inclusion de nouveaux magasins de données, pour par exemple prendre en compte les antécédents des patients dans les analyses. D'autres sorties que des rapports statistiques étaient également envisagées. Notre modèle multidimensionnel permet bien sûr une navigation OLAP, mais des vues « attributs-valeurs » pourraient également être extraites de notre entrepôt pour appliquer des techniques de fouille de données.

Le second type de perspectives était plus technique et visait à améliorer notre prototype. Cela incluait l'automatisation et la généralisation du processus d'ETL (*Extract, Transform, Load* – extraction, transformation et chargement) à tous les magasins de données, ainsi que l'amélioration de la convivialité et de la sécurité globales du système, cette dernière étant évidemment primordiale dans le cadre du traitement de données médicales personnelles.

Remerciements

L'auteur remercie le Docteur Jean-Marcel Ferret, médecin du sport et porteur du projet MAP au sein de l'incubateur d'entreprises Créalys. Ce travail a été cofinancé par la Région Rhône-Alpes et l'Université Lumière Lyon 2.

5. Bibliographie

Ben Messaoud, R., Rabaseda, S., Boussaïd, O., Bentayeb, F., « OpAC: A New OLAP Operator Based on a Data Mining Method », *Proceedings of the 6th International Baltic Conference on Databases and Information Systems (DB&IS 04)*, Riga, Latvia, 2004.

Firestone, J.M., « DKMA and the Data Warehouse Bus Architecture », *DKMS Brief*, n° 7, 2002, http://www.dkms.com/papers/dkmadwbus.pdf.

Inmon, W.H., *Building the Data Warehouse*, John Wiley & Sons, Hoboken, 2002.

Kimball, R., Ross, M., *The Data Warehouse Toolkit*, John Wiley & Sons, Hoboken, 2002.

Saad, K., « Information-based Medicine: A New Era in Patient Care », *Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP (DOLAP 04)*, Washington DC, USA, 2004, ACM Press, New York, p. 58.