

医院数据仓库数据模型设计

汪涛

(合肥工业大学 计算机与信息学院, 安徽 合肥 230009)

摘 要:数据模型设计是数据仓库建设的核心,提出一种医院数据仓库数据模型的设计方法。以某一三甲医院的 HIS 数据为背景,采用数据驱动的手段,结合医院的需求,提出了医院数据仓库的三层数据模型,概念模型、逻辑模型、物理模型,并完整地给出了每个模型的具体设计和主要内容。设计并实现了医院数据仓库的数据模型,并结合医院具体的数据给出了相应的实例。此医院数据仓库的三层数据模型易于理解和实现,为医院数据仓库设计最终完成提供了基础。

关键词:数据模型;概念模型;逻辑模型;数据仓库

中图分类号:TP39

文献标识码:A

文章编号:1673-629X(2010)05-0191-04

Data Model Design of Hospital Data Warehouse

WANG Tao

(School of Computer & Information, Hefei University of Technology, Hefei 230009, China)

Abstract: Data model design is the core of building a data warehouse, present a kind of design approach of hospital data warehouse's data model. With a top three hospital HIS data as the background, using data-driven tools and combined with the needs of the hospital, the text proposes the hospital data warehouse's three-tier data model of conceptual model, logical model, physical model, and completes each model given the specific design and main content. This hospital data warehouse's three-tier data model is easy to understand and realize and provides the basis for the design finalized.

Key words: data model; conceptual model; logical data model; data warehouse

0 引言

随着医疗市场的竞争越来越激烈,为了提高医院的竞争力,各家医院对信息化建设投入不断加大^[1]。医院信息系统的使用提高了基本业务处理的效率,提升了管理的手段。但随着时间的推移,现有系统积累了海量数据,如何对其中的各类业务数据加以整合和利用,从中挖掘出隐藏在背后的有价值、可以利用的潜在信息,对以后医院科学的业务分析和决策管理具有十分重要的意义。

数据仓库的出现正好可以解决以上问题,如“军字一号”医院信息系统上建立数据仓库,整合和分析历史数据^[2],为医院决策提供数据。而数据仓库系统如何对海量数据进行有效组织和管理,并使之支持千变万化的管理业务分析与决策,主要依赖于数据仓库系统逻辑数据模型(Logical Data Model,简称 LDM)的设计^[3]。

一个好的逻辑数据模型能够最大地保证灵活性和

可扩展性,以满足数据源的变化和应用需求的拓展。因此建设好 LDM 是医院数据仓库的关键,文中就此进行探讨。

1 数据仓库数据模型的概述

数据仓库是一个面向主题的、集成的、非易失的且随时间变化的数据集,用来支持管理人员的决策,其与管理型数据库系统(OLTP)的建模方法是不同的^[4]。操作型数据库系统是为具体的业务活动,是在传统开发生命周期(SDLC)下进行的,但不适用于决策支持系统领域。

在用户需求尚不明确的情况下,数据仓库的建模是从整合现有操作型数据开始的,分析业务系统的数据组织、关系模型,确定数据范围、主题,依次设计系统的概念模型、逻辑模型和物理模型。

而在实际的项目实施当中,如在医院的数据仓库建设中,系统的初步需求还是需要首先了解和分析的。这为数据仓库建设提供了原始范围,以及最终为用户所接受提供保证。以下,笔者在设计医院数据仓库模型中是从系统需求分析和 HIS 的数据分析开始的,采用三层模型设计的。

2 设计医院的数据仓库的数据模型

2.1 概念模型的设计

2.1.1 系统边界的确立,包括需求分析、数据来源等

现有的医院数据是面向具体业务的,缺少数据分析,难以为领导的决策服务;数据仓库提供了对现有、历史数据的分析,为医院决策者感兴趣的问题找到答案^[5]门诊、住院病人数量变化趋势怎样?与季节、医生有何关系?药品比例过高的科室哪些医生开方最多?哪种药品最多?病人治疗效果分布如何及其关联因素?收入增长当中有多少是来自非药品?

我院数据仓库的来源主要是现有 HIS(医院信息系统)积累的数据和医院病案首页数据。包括门诊病人就诊信息费用信息、住院病人费用信息,住院病人的分布信息、诊疗信息,药品采购、使用的动态信息,各种医保病人结算及相关信息。

2.1.2 确定主题及其内容

主题是在较高层次上将企业业务模型和面向事务的数据进行分析、归类和综合的一个过程,每个主题对应了一个分析领域。数据仓库中的数据是面向主题的,主题的确立过程主要的工作包括对医院的业务模型的分析,信息系统数据的 E-R 图的分析,以及医院宏观决策需求分析。现根据需求分析和现有数据的综合,可以确定如下四个主题:

费用主题(分为门诊和住院),其主要内容有病人 ID、处方 ID、科室、医生、处方、费别、病人类型、时间、费用金额;

药品主题(采购和使用),采购部分主要内容有采购的药品名称、药房、时间、供应商、采购人员、采购价格,使用部分主要内容有使用药房、药品名称、时间、处方、医生、科室;

住院病人的分布,这部分数据主要来自病案首页的数据,主要内容有姓名、性别、年龄、职业、入院时间、入院诊断;

病种分析,主要内容有病人 ID、疾病诊断名称、治疗效果、药物名称、手术名称、科室、主治医师、治疗总费用,住院日。

2.1.3 建立数据仓库概念模型

概念模型设计的目的是在需求分析和主题分析的基础上建立一个较为稳固的概念模型。在这里采用信息包图完整、规范化地分析上述主题,同时建立各个主题的星型图。

信息包图建立方法如下:

①、通过对需求分析的结果的整理,提取医院领导和各部门主要领导所关心的主要指标。如:院长关心每月的收入变化趋势,医务处关心手术次数及医疗相

关的变化趋势,药剂科主任关心药品采购、销售、使用的变化。

②、分析相关指标的分析角度,即未来要设计星型模型的维度。如费用主题中,当领导需要了解医疗收入这一指标时,他们分析的角度一般上是时间周期、科室、费别、病人类型、医生。

③、对维度进行分析,确定每个维度的层次结构,确立多维数据分析的上钻、下钻的途径。如在费用分析里,时间周期级别有年、月和日,不需要更小的层次了,费别中领导很关心药品比,所以增加了高一级的药品和非药品。

查看现有 HIS 中的数据,其中费用信息最为详细和完整,现以费用主题、住院病人作为对象,建立信息包图,如图 1 所示。

层次结构	维度				
	时间周期	所在科室	医生姓名	费别	病人类型
	年	科室名称	医生姓名	药品和非药品	是否医保
	季			各类别	各类别
	月				
	日				
度量值: 收入(费用)					

图 1 费用信息包图

信息包图中可以确定该主题分析的目标是收入的变化分析,可以从年、季度、月甚至日的角度,从科室、医生、费别、医保类别来研究费用的变化。

采用星型架构的方法来表示这种数据关系,星型架构由一个事实表和一组维表组成。事实表的主键是由各个维表的主键共同组成,而非主属性是数值性数据,也是要考察的指标数据。此主题的星型架构图见图 2。

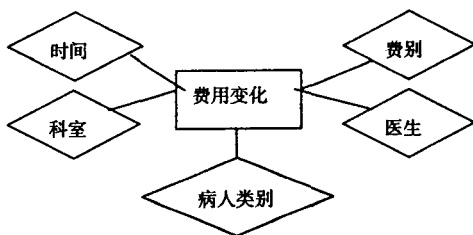


图 2 星型架构的逻辑视图

2.2 逻辑模型的设计

2.2.1 详细分析主题域

上面确定了几个基本的主题域,在进行设计时,一般是一次一个主题或一次若干个主题地逐步完成的。所以,必须对概念模型设计步骤中确定的几个基本主题域进行分析,一并选择首先要实施的主题域^[6]。

我院 HIS 是 1998 年开始上线,主要是门诊、住院

收费,2003年开始了医护工作站的使用,很显然这些数据都是以财务收费为主体的,所以费用信息在我院历史数据中最为详细。同时医院领导最为关心的也是收入的变化,以及药品比例的变化。所以,文中首先分析住院费用主题,逐步设计数据仓库的完整数据模型。

分析费用主题的详细属性,从 HIS 数据库当中提取、集成数据导入到数据仓库,作为多维数据分析的基础数据源。费用主题的主属性有:科室 ID、医生 ID、发票 ID、费用小计、病人 ID、时间。源数据的相关表有:

病人信息表:病人 ID、病人姓名、性别、年龄、地址、病人类别等;

处方信息表:病人 ID、处方 ID、科室名称、医生姓名、项目名称、数量、单价、小计、发票类别、时间。

还有发票类别表与病人类别表。根据这些源数据下面就可以设计该主题的事实、维度表了。

2.2.2 粒度层次划分

粒度问题是设计数据仓库的一个最重要方面。粒度是指数据仓库的数据单位中保存数据的细化或综合程度的级别。它深深地影响存放在数据仓库中的数据量的大小,同时影响数据仓库所能回答的查询类型^[4]。因此必须在数据仓库中的数据量大小与查询的详细程度之间作出权衡。

对于费用变化,领导一般关心的月报、季报、年报,而无须关注日数据的变化;同时也是以科室为单元,区分药品、非药品、院内制剂、手术费用的比重。因此可以考虑采用双重粒度,分别建立一个具有上述属性细节数据库,确保了源数据在数据仓库中细节不会丢失,再建立一个领导经常使用的轻度综合的数据库,见图 3。

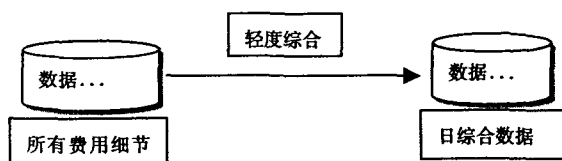


图3 双层粒度

以 2007-2008 年的数据为例,对于细节库中费用明细表的设计,首先要考虑到历史病人费用清单的查询,通过增加冗余,把病人的姓名、性别、病人类别等字段加入到处方信息表,从而构成了一个费用细节表,行数为 9 百多万条;而在汇总的数据库中,要进行汇总,只保留了病人 ID、发票 ID、科室 ID、医生 ID、日期(年月日)费用小计,此时行数为四百多万条,数量大大减少。通过采用双重粒度,把经常访问的数据存储在高速设备上,节约了空间,也提高了访问的效率。

2.2.3 确定分割策略

数据分割的目的是把数据分成小的物理单元,便于数据的处理和灵活的访问;数据分割的表准是根据设计需要灵活的划分的^[4]。医院的数据主要是病人的信息,以时间为顺序的,跨年度的病人比例一般比较小,而且费用汇总也是以时间为单位的,所以数据分割以时间为标准,可以形成每年的数据库,便于数据的查询、分析和综合。文中以年分别建立数据库,作为数据仓库的源数据。

2.2.4 确定星型架构的构成:事实表和维度

星型架构主要由事实表和维度表构成,这种模型的中心是事实表,包含了度量值和维度表的主键,维度表主要存储了各个维的具体字段,事实表和维度表之间是通过主键和外键关系相联系的。在设计该模型当中,层次结构的设计是非常重要的,层次结构就是用户在分析数据时向上向下钻取的路径。以时间层次机构为例,一般上年是最高层次,接下来依次是月、日,那么时间维度必须包含这几个字段,并严格约束它们之间的对应关系。在下面费用分析中,用户就可以很轻松地了解到一年中每个月、一个月中的每一天的收入金额。另外为了药品比的分析可在类别(RevoiceID)维当中加入药品判断字段,层次设计放到较高级别就可以了。

图 2 已经给出了星型架构的逻辑视图,根据对应的主题分析,就可以设计相应的星型模型^[7]。主题有费用主题、药品主题、病人主题、病种主题,都可以设计出自己的事实表和维度。在设计维度的时候,可以几个主题共享同一维度,可以确定时间维度、病人维度可以共享。这里还是以费用为主题设计它的星型模型,见图 4。

费用事实表、维表的设计是服从于医院费用分析。药品比、医保构成比是比较关心的问题,因此在费用类别里添加了药品、非药品字段,病人信息表里添加了是否医保字段,这样就可以建立包含药品和非药品发票类别的层次结构以及包含医保和非医保病人类别层次结构了。

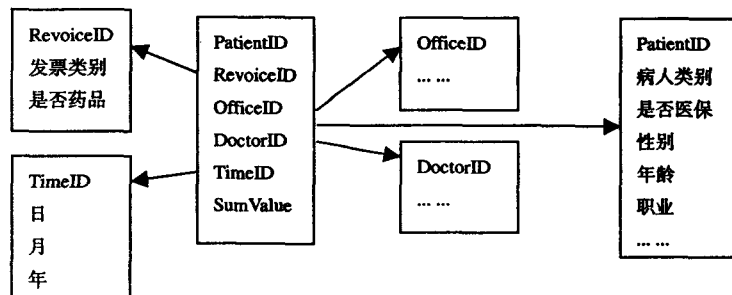


图4 星型模型图

同时根据粒度的分析,时间维最小分析单元是日,就可以把源数据时间格式转换成标准的日期格式(YY-MM-DD),按病人 ID、科室名称、医生姓名、发票类别、日来汇总小计的金额。这样大大减少了记录条数,在对医院 HIS 的 2007-2008 年数据处理时,可以减少 50% 的记录。

3 物理模型的设计

数据仓库的物理模型的设计就是数据仓库逻辑模型在物理系统的实现。其中主要解决数据的结构类型、数据的索引策略、数据的存储策略及存储优化等问题。在进行物理模型的设计实现时,所考虑的因素有:性能价格比、I/O 存储时间、空间利用率及维护的代价^[8]。

以我院为例,每天产生的数据处方细节表就有几万条,细节数据量积累起来很快。物理设计时,把 OLAP 中使用频率高的、要求响应时间快的放在高速存储设备中。对于细节性数据、访问次数较少的放在低速、价格低廉的介质上。这样也可以提高高速存储设备的利用率,降低成本提高了性能价格比。

数据仓库数据量很大,建立索引会大大地提高访问效率。建立索引是要考虑空间成本的,并不是所有的表都需要索引,只需对访问比较频繁的表建立索引,同时根据时间需要选择索引的键。例如,在费用的关系表中,时间维度访问频率最高的,可以在事实表中建立 TimeID 的索引。

在设计存储策略中可以有许多方法提高 I/O 的效率,此处介绍这次数据仓库中用到的方法。

①增加冗余,将经常用到的字段添加到相应的表里,减少表的连接次数以及 I/O 的访问次数,提高访问速度。如,在开始设计病人维度时,病人类别是一个单独的关系表,其字段并没有放在病人维时,这样在分析费用中需要做两表的连接,降低了访问效率。

②归并表,即将需要同时访问的表顺序存放在同一个物理磁盘上,减少 I/O 访问的次数。

4 结束语

数据仓库的数据模型设计是数据仓库建设的核心,模型设计的好坏决定了数据仓库项目成功完成与否。文中以医院具体数据为背景讨论了概念模型、逻辑模型、物理模型的设计方法和主要内容。

医院建设数据仓库的数据源是 HIS,数据范围在不断增加,数据字典时有变动,用户需求也随着不断变化,数据模型的设计本身就是个不断循环和修正的过程^[9]。因此,医院数据仓库模型的设计也要遵循循环的开发方法,在用户看到实际结果后,又有可能提出更完善或新的需求,通过反馈使得数据模型不断的修正,最后形成适合医院的数据仓库模型。

参考文献:

- [1] 段会龙,吕旭东. 医疗信息系统发展现状及趋势[J]. 中国医疗器械信息,2004(10):1-6.
- [2] 毛琦敏. 数据仓库在医院应用的研究[J]. 医学研究生学报,2005,18(4):358-359.
- [3] 胡黎玮,苏宁军. 浅谈数据仓库成功实施的关键因素[J]. 科技情报开发与经济,2009,19(5):81-83.
- [4] Inmon W H. 数据仓库[M]. 王海志,等译. 北京:机械工业出版社,2006:27-58.
- [5] Wisniewski M F, Kieszkowski P, Zagorski B M, et al. Development of a clinical data warehouse for hospital infection control[J]. Journal of the American Medical Informatics Association, 2003,10(5):455-461.
- [6] 谷岩,郭庆. 数据仓库系统中逻辑建模的方法研究[J]. 计算机系统应用,2005(8):42-46.
- [7] Gyssens M, Lakshmanan L V S. A foundation for multi-dimensional databases[C]//Proceedings of the 23rd International Conference on Very Large DataBases. Athens, Greece: [s. n.], 1997:107-114.
- [8] 朱德利. SQL Server 2005 数据挖掘与商业智能完全解决方案[M]. 北京:电子工业出版社,2007:123-126.
- [9] Mallach E G. Decision Support and Data Warehouse Systems[M]. Beijing: Tsinghua University Press, 2001:201-236.

(上接第 190 页)

- [J]. 计算机工程与应用,2008,44(11):43-46.
- [6] Thomas P R, Xin Y. Stochastic Ranking for Constrained Evolutionary Optimization[J]. IEEE Trans. on Evolutionary Computation, 2000,4(3):284-294.
- [7] Mezura M E, Coello C A, Tun M I. Simple Feasibility rules and Differential Evolution for Constrained Optimization[C]//Proceedings of the 3rd Mexican International Conference on Artificial Intelligence, LNCS 2972. Heidelberg: Springer-Verlag, 2004:707-716.
- [8] Mezura M E, Carlos A, Coello C A. Simple Evolution Strategy to Solve Constrained Optimization Problems[J]. IEEE Trans. on Evolutionary Computation, 2005,9(1):1-17.
- [9] 熊敏,刘玉树. 基于协同进化遗传算法的地域选取算法[J]. 计算机技术与发展,2006,16(6):174-176.
- [10] Koziel S, Michalewicz Z. Evolutionary Algorithms, Homomorphous Mappings, and Constrained Parameter Optimization[J]. Evolutionary Computation, 1999,7(1):19-44.