广西科学院学报
Journal of Guangxi Academy of Sciences
2010,26(4):446~448,454
Vol. 26, No. 4 November 2010

一种基于 ER 模型的数据仓库多维建模方法

A Multi-dimensional Modeling Mechanism of Data Warehouse Based on ER Model

刘洪磊[1],李卫玲[2],王锡民[1],余红兵[1]
LIU Hong-lei[1], LI Wei-ling[2], WANG Xi-min[1], YU Hong-bing[1]

(1. 96251 部队司令部,河南洛阳 471003;2. 洛阳师范学院信息技术学院,河南洛阳 471022)
(1. The Headquarters of 96251 Troops, Luoyang, Henan, 471003, China; 2. College of Information Technology, Luoyang Normal University, Luoyang, Henan, 471022, China)

摘要:结合 Ralph Kimball 和 Inmon 的数据仓库设计和架构思想,以建设某学院办公自动化数据仓库为例,提出一种基于实体联系(ER)模型的数据仓库多维建模方法。该方法从构建数据仓库全局角度指导维度建模,同时兼顾用户需求与底层数据,从而得到既能满足用户需求、又具扩展性的维度模型。该方法规范了数据仓库的逻辑模型建模过程,可以为多维模型的设计提供方法上的指导。

关键词:数据仓库 ER 模型 多维建模 办公自动化
中图法分类号:TP311 **文献标识码**:A **文章编号**:1002-7378(2010)04-0446-03

Abstract: Based on Ralph Kimball and Inmon's design methods and the architecture idea of Data Warehouse, a multi-dimensional modeling mechanism is proposed based on ER (Entity Relationship) model. A data warehouse of college Office Automation is built as an example. The method provides dimensional modeling guidance for building a data warehouse from a global point, simultaneously considering users needs and the underlying data. Thus an extended dimension model with the users satisfactory can be obtained. This mechanism standardiezes the logical modeling process of data warehouse, and provides methodological guidance for the design of multi-dimensional model.

Key words: data warehouse, ER model, multi-dimensional model, office automation

20 世纪 80 年代中期,"数据仓库之父" W. H. Inmon 定义了数据仓库的概念,随后又给出了更为精确的定义:数据仓库是一个面向主题的、集成的、非易失的、随时间变化的用来支持管理人员决策的数据集合[1]。这是目前世界公认的最为准确、全面的关于数据仓库的定义。数据仓库是面向企业的,建立数据仓库的工作范围和成本常常是巨大的,而且建设周期非常长,是高代价项目[2]。这对于中小型企业来说,往往是不可能接受的。因此,数据集市就应运而生。数据集市可以认为是一种更小的、更集中的数据仓库,这为企业分析商业数据提供一种廉价途径。全世界对数据仓库总投资的一半以上均集中在数据集市上[3]。数据仓库的数据模式主要有 3 种结构模式:星型模式、雪花型模式和星座模式[4]。

建设数据仓库项目必然涉及到数据仓库建模的问题,合理而完备的数据模型是用户业务需求的体现,是数据仓库成败的技术因素。传统的联机事务处理(OLTP)系统是面向应用的,建立在实体联系(ER)数据模型的基础上。而数据仓库是面向主题的,建立在多维模型的基础之上。如果能够根据传统数据库的 ER 模型转换得到数据仓库所需的多维模型,充分利用原有 ER 模型隐含的信息,在很大程度上可以缩短数据仓库系统的开发周期。

Ralph Kimball 和 Inmon 都是"数据仓库"的首创者,但是在数据仓库的设计和架构思想上却很不

收稿日期:2010-08-11
修回日期:2010-10-13
作者简介:刘洪磊(1979-),男,硕士,主要从事数据仓库研究。

相同。Kimball[5]认为通过描述使用一致性维度,能够从建设企业级数据仓库全局角度分析不同数据集市中的信息,即可以通过一系列维数相同的数据集市递增地构建数据仓库。但是其缺点是结构化程度不高;数据仓库的构建从需求出发而不是从底层数据出发来设计。Inmon[6]主张将不同的 OLTP 数据集中到面向主题、集成的、不易失的和时间变化的结构中,用于以后的分析;并且数据可以通过下钻到最细层,或者上卷到汇总层;每个数据集市是针对独立部门特殊设计的。其思想是能够保证一种更加结构化的方式开发维度模型,并且保证数据仓库模型反映底层数据关系。缺点是对需求考虑不足,构建的数据集市缺乏全局性考虑,不利于整体协调构建数据仓库。可见上述两种建模方法都各有优势与缺陷,但是同时它们之间也具有一定的互补性。本文结合 Ralph Kimball 和 Inmon 的数据仓库设计和架构思想,以建设某学院办公自动化数据仓库为例,提出一种基于 ER 模型的多维建模方法。

1 源系统概述

某学院办公自动化系统是采用目前较为流行的 Browser/Server 体系结构、JSP 技术以及 Struts 框架而建立的综合信息管理系统。目前,关系型数据库中已经积累了大量的教学、科研、人事、学科建设等数据信息,形成了比较完整的数据。这些数据真实地放映了该学院的各项基本情况,可供学院内部进行各类统计报表和信息查询。

用户从原业务数据和一些相关的外部数据(如国家的相关政策、规定等)中通过 ETL 完成数据的抽取、转换、清洗,并分别装载到本科生教学、研究生教学、本科生日常、研究生日常、学科建设、科研管理、人事管理、业绩管理等八个数据集市中。

如图 1 所示是源系统的科研管理业务模型中与项目进款相关部分的 ER 图举例,其他业务模型 ER 图与其相似。在本例中 Amoney 实体为进款实体,与项目基本信息实体 project、时间实体 time、教师分配款项信息实体 Assign 之间都是多对一的关系,并且教师分配款项信息实体 Assign 与教师信息实体 teacher 是多对多的关系,含义为某个教师的某个项目在某天有一笔进款,进款金额由 arrive_number 字段记录。项目基本信息实体 project 和项目来源实体 source 是多对一的关系,而项目来源实体 source 和项目来源类型实体之间也是多对一的关系,即某一项目来源类型对应多个项目来源,多个项目属于同一项目来源。同时教师基本信息实体 teacher 和岗位实体 post、职称实体 duty、单位实体 unit 之间具有一对多的关系。

2 建立数据仓库总线结构矩阵

首先构建系统数据仓库的总线结构矩阵以获得一个整体的视图,并确定一致性维度,从而可以使我们从构建数据仓库全局角度来对维度模型进行设计。如表 1 中学科、学生、论文、教师、时间等维度在多个数据集市中出现,可以设计为一致性维度。其次,从选取的数据集市相关的业务数据模型中导出维度模型。最后,必须充分考虑用户需求。

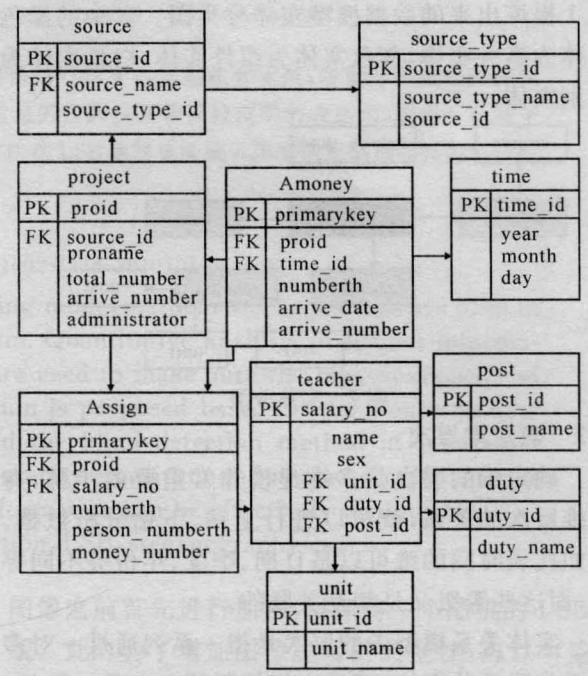

图 1 源系统的科研管理业务模型

表 1 系统的数据仓库总线结构矩阵

	时间	教师	专利	著作	论文	项目	学生	基地	学科	教学立项	教学成果	就业	奖学金
本科生教学	◇						◇			◇	◇		
研究生教学	◇	◇			◇		◇		◇				
本科生日常	◇						◇					◇	◇
研究生日常	◇						◇					◇	◇
学科建设	◇	◇	◇	◇	◇	◇		◇	◇				
科研管理	◇	◇			◇	◇		◇	◇				
人事管理	◇	◇			◇	◇		◇	◇				
业绩考核	◇	◇							◇				

3 由 ER 模型创建多维模型

3.1 实体分类

多维建模的首要任务是确定事实数据和维数据。首先应对 ER 模型中的实体分类,进而识别事实与维数据。从构建多维数据模型角度出发,实体

可分为事务实体、组件实体和级别实体。

第一类是事务实体。事务实体记录了事件发生的操作细节信息，它们构成了星型模式中事实表的基础。

第二类是组件实体。组件实体是指通过一对多关系与事务实体关联的实体，组件实体是构成多维数据模型中维度表的基础。

第三类是级别实体。级别实体是通过一对多关系与组件实体相关联的实体，它们都直接或者传递性地依赖于组件实体。级别实体描述了数据模型中的层次关系，它们可以合并进组件实体中，从而形成多维数据模型的维表。如图2所示的实体分类是从图1提炼出来的数据模型实体分类图。其中的黑色实体为事务实体，灰色实体为组件实体，白色实体为级别实体。

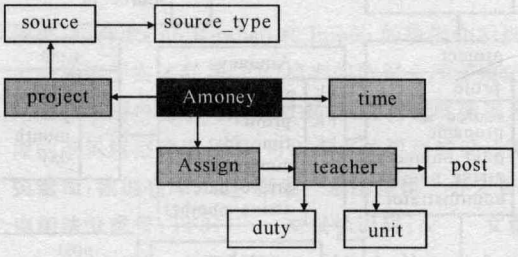

图2 实体分类

3.2 确定维层次

确定维的层次是多维建模非常重要的步骤，有了维层次的关系，就可以进行上钻、下钻分析数据。例如代表时间的维可以是日期、季度、年份等不同等级，而这些等级又是相互关联的。

实体关系模型中的层次是指一系列通过一对多关系相联系的实体，而且它们相联的方向是一致的。如图3所示，unit - teacher - Assign - Amoney就是一个层次。

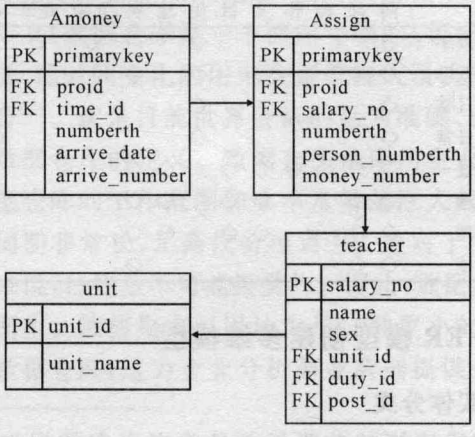

图3 层次举例

3.3 生成维度模型

在完成了实体分类和确定维层次之后，还需要再通过合并实体和汇总数据两个操作才能生成最终的维度模型。

3.3.1 合并实体

合并实体即将两个或更多的实体合并成一个。合并实体通过减少连接操作的数量，提高数据交付处理的性能，并且可以增加一致性。例如，我们可以将teacher实体;post实体、unit实体和duty实体合并。此时，teacher实体中的属性就是它原本的属性加上合并进来的unit_name、duty_name和post_name属性。

3.3.2 汇总数据

汇总数据是将一个事务实体通过汇总操作转换成新的包含概要数据的实体，最常见的汇总标准是时间。汇总数据的目的是使得在线存储的需求减少、分析标准化及改善数据的交付性能。

合并维表：有时可以将相关的维表合并成一个维表，以提高使用的性能。

合并事实表：为了减少最终的维度模型中的星型模式的数量，应当将具有相同的主键的事实表进行合并。这样可以使原来在不同事实表中的度量处于同一个事实表中，以利于度量值之间的比较。

通过上述步骤，最终得到的星型维度模型如图4所示。

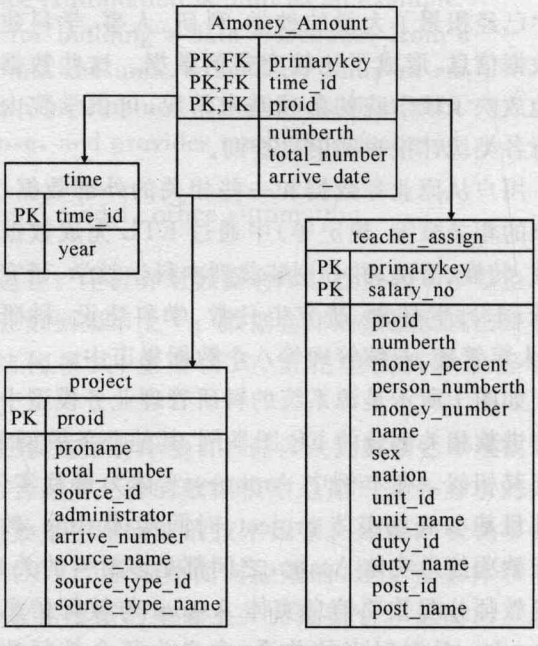

图4 星型维度模型

4 结束语

本文以某学院办公自动化数据仓库建设为例，提出了一种基于ER模型的多维数据建模方法。应

（下转第454页）

```
sh. Run "telnet "+ipadd
WScript. Sleep(1500)
sh. appactivate  "telnet "+ipadd
sh. SendKeys "jiaohuanjimima{ENTER}"
WScript. Sleep(1500)
sh. appactivate  "telnet "+ipadd
sh. SendKeys "en{ENTER}"
WScript. Sleep(1500)
sh. appactivate  "telnet "+ipadd
sh. SendKeys "jiaohuanjimima{ENTER}"
WScript. Sleep(1500)
sh. appactivate  "telnet "+ipadd
sh. SendKeys "copy flash:config. text tftp://
10. 1. 14. 6/"+name+". txt{ENTER}"
WScript. Sleep(2000)
sh. appactivate  "telnet "+ipadd
sh. SendKeys "exit{ENTER}"
sh. appactivate  "telnet "+ipadd
sh. SendKeys "a"
wend
f. close
```

针对交换机的现场调试问题，关键是解决计算机如何通过Trunk口接入网络。事实上，只要有支持802.1q的网卡，就能解决此问题。目前大部分的笔记本计算机网卡都支持802.1q，但是在Windows环境下缺少驱动，不能输出带VID标记的数据帧，而在Linux下却能轻易实现。新版本的Linux支持VLAN功能，只需在终端上输入ifconfig add eth0 glvid，就能在eth0上增加VID为glvid的网卡子接口。设定glvid为管理网段VID、子接口的IP地址为管理网段IP地址和相应的网络参数后，就能通过Trunk口访问交换机和网络，进行设备调试。对于采用802.1x实施接入认证的网络，Linux下也有相应的认证软件，实际使用时，可以将Linux系统制作成虚拟机，调试时再启动。

3 结束语

VLAN技术能提高网络的传输效率和安全性，在局域网中得到了广泛应用。在VLAN环境中交换设备的调试仍然采用串口调试、Web调试和Telnet程序3种方法进行，但是存在网络交换设备多、配置管理复杂，以及交换机现场调试不方便两个问题，本文提出通过vbs脚本结合Telnet和Tftp程序对数量众多的交换设备配置进行批量管理，利用Linux解决Trunk端口引发的交换机现场管理问题，从而保障网络顺利运行。

参考文献：

[1] 沈海娟. 网络互联技术——路由与交换[M]. 杭州：浙江大学出版社，2006.

[2] 黄彪. VLAN中Tagged与Untagged探讨[J]. 电脑知识与技术(学术交流)，2007(2)：377-378.

[3] 周昌权. 局域网端口VLAN技术的实现[J]. 电脑编程技巧与维护，2010(2)：96-97.

(责任编辑：韦廷宗)

(上接第448页)

用证明，本文所采用的方法从构建数据仓库全局角度指导维度建模，同时兼顾用户需求与底层数据，从而可以得到既能满足用户需求的、又具扩展性的维度模型。同时，该方法通过提供一组全面的重构操作，规范了数据仓库的逻辑模型建模过程，缩短了数据仓库系统的开发周期。

参考文献：

[1] Inmon W H. Building the data warehouse [M]. New York：John Wiley & Sons Inc，1993.

[2] 梁洁敏. 银行数据仓库系统的设计与实现[D]. 济南：山东大学，2005.

[3] 计算机世界网. 数据仓库[EB/OL]. [2009-08-14]. http://www2.ccw.com.cn/1997/10/155322.shtml.

[4] 于戈，王大玲，鲍玉斌. 数据仓库设计[M]. 北京：机械工业出版社，2004.

[5] Ralph Kimball. 数据仓库生命周期工具箱[M]. 第二版. 唐富年，孙媛媛，译. 北京：清华大学出版社，2009.

[6] Inmon W H. 数据仓库[M]. 王志海，林友芳，译. 北京：机械工业出社，2000.

(责任编辑：韦廷宗)