中国农业大学学报 2003,8(4):44～46
Journal of China Agricultural University

构建数据仓库模型的一种方法

彭 波 余顺芝
(中国农业大学 信息与电气工程学院,北京 100083)

摘 要 构建了具有 4 个存储结构和 3 个数据流的数据仓库模型,并以参与设计的数据仓库系统为例,论述了数据仓库模型的构建过程。较之其他数据仓库系统,本系统创建了第二索引,即一个高性能的存储,以空间换时间,既实现了企业数据集成,又提高了系统的查询速度;在数据仓库中设置了代理键,简化了多数据源数据的集成工作,提高了数据抽取、转换和加载的速度,并为数据仓库系统的迭代式开发提供了有利条件。提出了构建数据仓库模型的建议:采用星系模型,原子级数据模型和汇总级数据模型并存的形式,设立代理键。

关键词 数据仓库;数据模型;数据存储
中图分类号 TP 301.6;TP 311 **文章编号** 1007-4333(2003)04-0044-03 **文献标识码** A

Study on building data warehouse model

Peng Bo, Yu Shunzhi
(College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China)

Abstract Being with taking the background of data warehouse for instance, the process of constructing the data warehouse model was discussed in detail, and a kind of data warehouse architecture containing four store-structures and three data-flows was put forward. Two new ideas was given compared with other similar systems, firstly, the second index for data warehouse was created, for the purpose of accomplishing the enterprise data integration and system query speed, which win the time at the cost of space. Secondly, the surrogate key was set in the data warehouse, which makes the integration of multiple data source simple, improves the speed of data draw-out, data conversion, data load and makes it possible for the iterative development to the data warehouse system. An advice of constructing the data warehouse model was put forward: adopting star model, merging atom data model and statistics data model, seting the surrogate key.

Key words data warehouse; data model; data storage

20 世纪 90 年代以来,国内电信业务发展异常迅速,其用户数量、业务范围也随之迅速增长和扩大。与此同时,行业信息化进程发展迅速,各大运营系统相继投入使用,积累了大量的历史数据,但是,就目前看来,这些数据在原有系统中无法提炼并升华为有用信息,不能及时提供给业务分析人员和管理决策者[1],经理人员被淹没在“数据的海洋”中[2];因此,企业需要一个能够容纳各种格式内部数据和外部数据,能够给企业决策者提供决策信息的系统[3],即数据仓库系统。

20 世纪 80 年代中期,数据仓库概念首次出现在“数据仓库之父”William H. Inmon 的《建立数据仓库》一书中[4],其定义为:数据仓库是支持管理决策过程的,面向主题的、集成的、随时间可变的、持久的数据集合①。简单来说,数据仓库是一个环境,而不是一件产品;是把事务型数据库中的操作数据汇总、集成之后产生的信息数据单独放到一起所形成的数据集合[5]。其主要用途是使用户更快、更方便地查询所需要的信息,为企业提供决策支持[6]。

收稿日期:2002-10-22
作者简介:彭 波,硕士,副教授,主要研究方向为计算机应用
① Ken Orr. Data Warehousing Technology. A white Paper, 1996

1　数据仓库体系结构

以笔者参与设计的江西联通数据仓库系统[7]为例,描述数据仓库模型的构建过程。该数据仓库系统由 4 个数据存储结构和 3 个数据流构成,其结构框图见图 1。

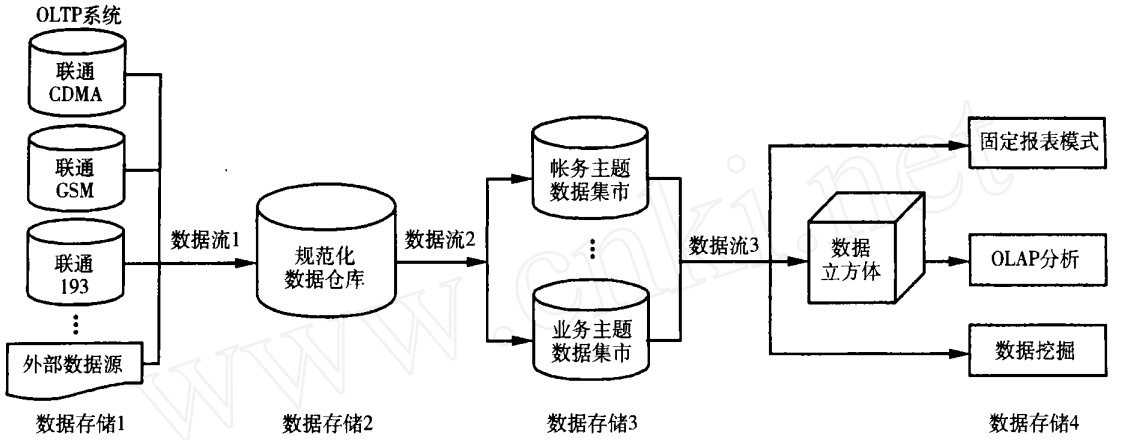

图 1　江西联通数据仓库系统体系结构

Fig. 1　Jiangxi 's union data warehouse system 's structure

1) 数据存储 1 ——OL TP 系统。数据存储 1 是联通公司原有的 OL TP 系统,为数据仓库系统提供业务数据来源。

2) 数据存储 2 ——数据仓库。又称为数据仓库系统的综合层,是一个规范化的数据库,存储着从原有 OL TP 系统中抽取并经过清洗和转化的干净数据。

3) 数据存储 3 ——数据集市。又称为高性能查询结构,是为了支持最终用户查询而专门设计的数据存储结构,存储经过汇总的统计数据。

4) 数据存储 4 ——最终用户手中的数据。这是一个供人们阅读的数据存储。

5) 数据流 1 ——从数据源到数据仓库。由于原有的 OL TP 系统为遗留系统,因此存储在 OL TP 系统中的数据不能自动导入数据仓库中。数据流 1 是一个从源系统抽取数据的抽取模块,实现数据从数据源到数据仓库的转移,包括抽取、转换和加载 3 部分。

6) 数据流 2 ——从数据仓库到数据集市。将数据从数据仓库加载到数据集市也需要一个数据加载模块,与数据流 1 一样,也有抽取、转换和加载 3 部分。

7) 数据流 3 ——从数据集市到前端应用程序。最终用户对数据的访问一般通过查询工具来进行,江西联通数据仓库系统采用 Brio 公司的 Brio Enterprise 实现数据展现功能。

2　数据仓库模型设计

2.1　需求确定

数据仓库系统的使用者是企业各级的决策人员,他们关心的问题和普通的营业操作人员不同[8],例如江西联通分公司的需求书中要求:查询年龄级别为 25～30 的普通男客户 2001 年 1 月 1 日的 CDMA 通话情况(包括通话次数、通话时长、市话通话时长、长途通话时长、市话费、长途话费 6 方面)。该问题涉及到以下信息:

1) 事实(Fact),即用于分析的基础数据。本文中为通话时长、市话通话时长、长途通话时长、市话费、长途话费。这类信息变化快,存储量大。

2) 维(Dimension),即分析问题的角度。从时间、客户、客户业务类型等几个角度分析通话情况。

3) 粒度(Grain),即划分维的单位。如时间维,可以按秒、分、小时统计;日期维可以按日、月、季度、年统计;客户维可以从客户的年龄、性别、客户级别来统计。由于维的粒度不同,所以模型设计可分 2 个阶段进行:第一阶段的原子级数据模型设计和第二阶段的概要表模型设计。

2.2　模型设计

数据仓库的建模技术已逐渐形成,并且在继续发展,目前比较流行的有 3 种:星型模型、雪花模型,以及结合星型模型和雪花模型优点的混合模型。本文中数据仓库设计采用混合模型的结构模式,其原子级数据模型示意图见图 2。该图中业务量事实表

属于星型模型事实表,周围的实体属于星型模型的维表,所示模型是数据仓库中有关客户通话情况主题的逻辑模型。将逻辑模型上升到物理模型后,业务量事实表包括 2 类数据:一类为表示分析主题的度量值,包括通话时长、市话通话时长、长途通话时长、市话费、长途话费;另一类为连接维表的外键,这些外键有一部分作为主键决定事实表的唯一性。

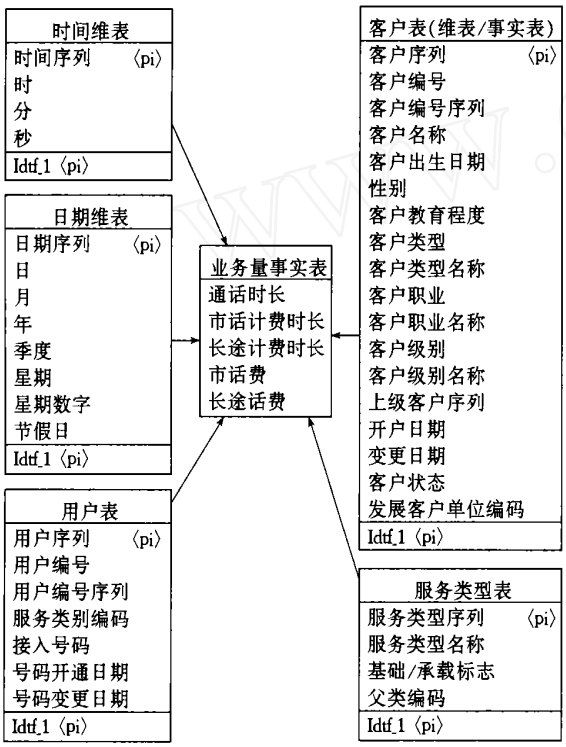

图 2 原子级星型数据模型示意图

Fig. 2 Atomic star schema data model map

原子级数据模型仅存储如某个手机号码在某一时刻的通话情况信息等粒度最低的数据。在实际应用中,所需数据多为经过汇总的,如 2.1 例中要求的男性 CDMA 用户在 2001 年 1 月 1 日全天的通话情况。这个用户有可能在这一天中打了多次电话,每次的通话情况都存储在原子级数据模型中。基于这种情况,需要建一个粒度高的概要表模型。图 3 是经过汇总的星型数据模型示意图,根据该模型的存储数据,可以查询诸如男性 CDMA 手机普通用户 2001 年 1 月 1 日全天的通话次数、通话时长等通话信息。概要表可以和原子级的数据表处在同一个数据仓库数据库中,也可以根据用户的需求,存放在独立的数据集市(数据存储 3)中。

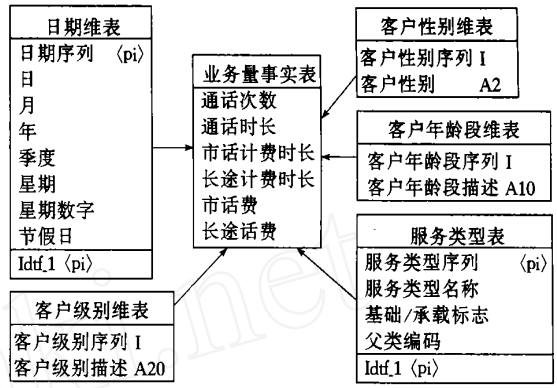

图 3 汇总级星型数据模型示意图

Fig. 3 Statistics star schema data model map

3 小结及建议

一个性能良好的数据仓库系统需要经过反复的试验和修改,笔者根据构建江西联通数据仓库系统的实际经验,提出了构建数据仓库模型的建议:

1)采用星系模型[9]。图 2 仅是针对一个需求的数据模型。在实际应用中,用户的需求多种多样,数据来源可能为多个事实表,故可采用多个事实表共存,之间通过公用的维表相关联的星系模型,也称为事实星座。

2)原子级数据模型和汇总级数据模型并存。坚持原子级数据模型和汇总级数据模型并存,而且要尽可能地细化原子级数据。因为用户的需求总是在不断增加,可以从细节数据中重建概要数据,但是不能在概要数据中建立细节数据;如果没有原子级数据模型,无形中丢失了许多重要的信息。

3)设立代理键(surrogate key)[10]。代理键是维表中一些没有业务含义的字段,只是一个由数据仓库加载程序时建立的数字。数据源表的主键仍然留在新建的数据仓库数据库表中,但不是新建表的主键。如客户维表中的客户编号和客户编号序列这 2 个字段,在 CDMA 系统中属于客户表的主键,但是在新建立的客户维表中仅作为一般的字段存在。

设立和使用代理键比业务键(business key)复杂得多,但是代理键提供了许多好处:

(a)获得维历史信息;

(b)提高查询数据的速度;

(c)便于从多个系统中整合同类信息;

(d)提高 ETL 速度;

(e)为数据仓库系统的迭带式开发提供条件。

(下转第 76 页)

分别建设城市生态农业休闲观光旅游带和绿化与特色经济林带,形成两带环抱三区、六园、九村的整体格局,促进城乡的可持续发展。

1.5 分区规划、支撑体系与保障措施

为了建设和组织管理需要,规划还从分区规划、技术支撑体系、基础设施建设、组织管理体制与运行机制创新、资源环境平衡和规划实施保障措施等方面进行专项研究[7]。

2 结束语

农业科技园区总体规划的主要思路是:凸显功能定位准确、优势产业突出、空间结构清晰、区域城乡联动的发展理念,重点构筑产业、技术和服务三大平台,吸引技术、人才、资金等多方位的投入要素和政府、企业、科教机构、村集体、农户等多元化的建设主体,聚集区内外各种资源,培育龙头企业,带动区域农业和农村的发展。规划研究的重点内容包括:建设目的和环境条件分析、规划总体思路和发展目标确定、园区功能和产业/产品市场分析定位、空间结构体系构建、功能分区和总体布局、分区规划和支撑保障体系的建立。笔者认为这一思路和研究内容对功能复杂的综合性农业科技园区的规划工作具有较强的针对性,在今后的研究中,将进一步完善和发展。对于特色和专业型园区的规划,应该针对各自特点,进行分类研究,提出相应的规划内容和方法,并以此为基础,提出能指导我国农业科技园区建设的规划理论和方法。

参 考 文 献

[1] 李学勇. 与时俱进创新进取开拓农业科技园区工作新局面[J]. 农村实用工程技术,2002(7):3～4

[2] 蒋和平. 中国农业科技园区的特点和类型分析[J]. 湖南农业大学学报(社科版),2000(6):14～17

[3] 沈悦林,徐四海,徐长明,等. 我国现代农业园区建设的动态和模式分析[J]. 农业现代化研究,1998(4):255～256

[4] 张宝文. 推进农业科技园区建设加速农业现代化进程[J]. 农村实用工程技术,2002(7):7～8

[5] 吴文良. 我国农业科技园区的发展定位与发展策略[J]. 中国农业科技导报,2001(3):18～19

[6] 孙振玉,贺晓丽. 试论农业科技园区建设的总体思路[J]. 农业技术经济,2001(4):21～22

[7] 卢凤君,孙世民. 长春农业科技园区建设和发展的战略思考[J]. 农业系统科学与综合研究,2002(2):142～145

[8] 王朝晖,李秋实. 农业高新技术产业示范区规划初探[J]. 城市规划,1998(4):25～28

(上接第 46 页)

参 考 文 献

[1] 柴 满. 数据仓库用于移动通信市场[J]. 微电脑世界周刊,1999(42):70

[2] Han Jiawei, Kamber M 著. 数据挖掘概念与技术[M]. 范 明,孟小峰,译. 北京:机械工业出版社,2001. 1～374

[3] 王建新,刘东波. 中国数据仓库应用市场等待激活[N]. 计算机世界,2001. 5. 31

[4] 张 澜,康增培. 数据仓库白皮书-概念篇[EB/OL]. 赛迪网,http://www.ccidnet.com,2001,3

[5] Inmon W H 著. 数据仓库(Building the Data Warehouse Ⅱ)[M]. 王志海,译. 北京:机械工业出版社,2000. 1～228

[6] 张 澜,康增培. 数据仓库企业的锦囊[EB/OL]. 赛迪网,http://www.ccidnet.com,2001. 2

[7] Corey M 著. Oracle8i 数据仓库[M]. 施平安,译. 北京:机械工业出版社,2001. 1～556

[8] 张忠能,尤 毅. 设计数据仓库[J]. 上海交通大学学报,1998,32(10):50～52

[9] Geiger G Jonathan. The data warehouse model[EB/OL]. http://www.dataWarehouse.com,2000. 9

[10] Demarest M. A data warehouse evaluation model[EB/OL]. http://www.hevanet.com,1995. 4