

论数据仓库中重要的数据模型

彭耶萍¹, 邓 婷²

(1. 吉首大学 信息工程系, 湖南 张家界 427000; 2. 湖南科技职业学院 机电系, 湖南 长沙 410008)

【摘 要】数据模型是数据库的基础, 良好的数据模型能为基于数据仓库的 OLAP 分析提供高效率和高质量的底层数据, 从而更有效地支持基于数据仓库的决策支持系统。

【关键词】OLAP; 雪花模型; 星型模型; 多维模型

【中图分类号】TP311.13 **【文献标识码】**A **【文章编号】**1673-1891(2007)03-0073-03

1 引言

基于数据仓库的 DSS (决策支持系统) 的迅速发展为决策者提供决策信息的底层数据的组织形式变得日益重要, 目前多数 DSS 是由数据仓库提供信息的, 在基于数据仓库的 DSS 的开发中, 数据模型是数据仓库中的一个重要问题, 是整个系统实现的基础。

2 星型模型

在关系数据库设计时, 采用实体关系模型 (E-R) 来描述事物之间的各种联系, 而在数据仓库中, 采用一个“事实表”和一些“维表”来描述事物多维的概念模型。事实表中的每个元组的指针指向了相应的一张维表, 例 1: 某公司对其产品作市场调查的角度包括销售时间、销售地区、订单这三个维度。如图 1:

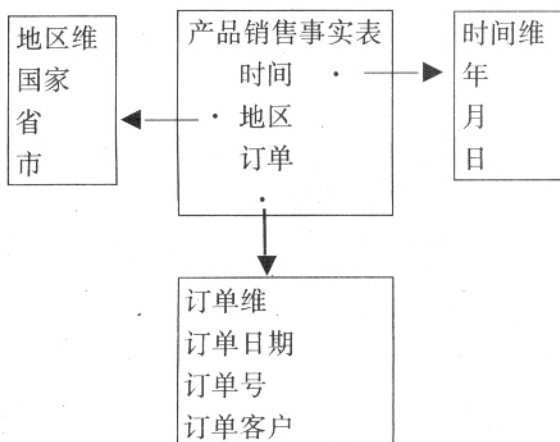


图 1 星型模型结构

星型结构中有多个维表, 但只有一个事实表, 整个星型结构对应一个事情的解决, 结构中的维表以事实为中心, 星型结构能明确描述一个事物的多维概念, 但在实际的应用过程中, 一个事实表往往超过了三个维度, 并且多个维表之间会产生维表间的共享, 例 2: 同例 1, 公司希望从不同的销售时间、地点、订单和客户的情况来调查市场, 并且要求能够查询出客户的基本情况。如图 2: 此时, 在这种情况下, 当问题涉及到的维度很多时, 星型结构会变得复杂, 数据组织冗余, 这样不易于元数据及系统的统一管理, 对数据仓库的多维操作速度会较慢。

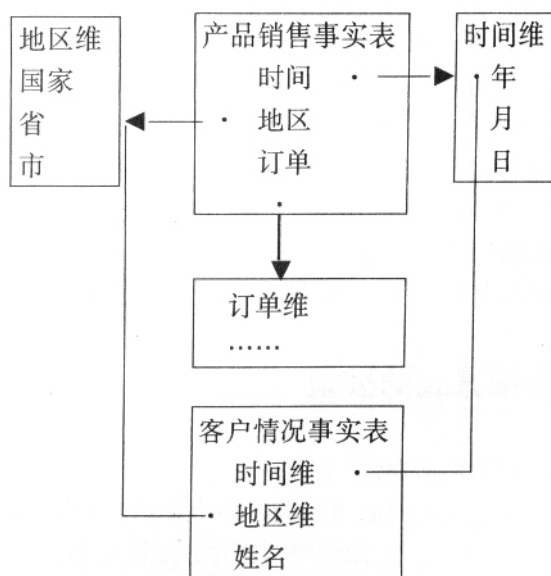


图 2 多维度的星型结构

3 雪花模型

收稿日期 2007-06-20

作者简介 彭耶萍 (1981-) 女, 湖南龙山人, 硕士研究生, 助教, 主要从事数据库研究。

雪花模型和星型模型都是基于关系数据库的,前者是后者的扩展。它可将星型中的某个角度的多维构成一个简单的“层次”,形成“雪花”结构,因此,在多维环境下雪花模型描述的事物多维会比较清晰。雪花模型是从根本上解决对数据表进行规范化处理,能够减少数据的冗余和数据的存储量,使系统更加优化。如下图:

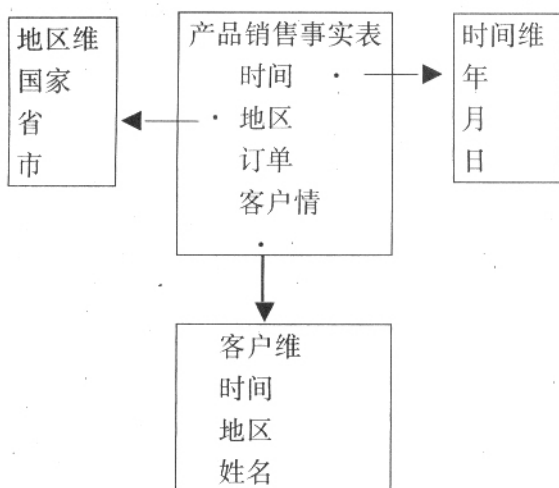


图 3 雪花数据模型结构

这样,事实表的数目将大大减少,雪花模型有利于程序设计人员和数据管理人员的接受,因此,目前多数数据仓库采用的是雪花模型,例 3: oracle 中,用 MOLAP 服务管理将多维操作影射成 SQL,再进行查询。但是雪花模型仍存在着不足之处,利用此模型在建立数据仓库时,会遇到大量的事实表之间的连接操作,这样既增加了事实表与相关维之间的 2 次连接运算,由使 OLAP (联机分析处理)的时间延长,如此会降低数据库的存储效率。因此,雪花型/规范化在 OLAP 模型中应用应当尽量限制。

4 多维方式的模型

4.1 OLAP 的多维性

Codd 提出 OLAP 标准,OLAP 操作具有多维的特征,关系数据库模型不能有效地描述及数据仓库的数据结构和语义,多维方式的模型是基于多维数组的数组,多维数组可以这样表示(维 1、维 2……维 N,变量),在各个维度的取值范围内进行标注,则可得到“数据立方体”,从而购成一个面向多维的数据库。例 4:从客户维,时间维,地区维反映公司某产品销售情况的三维数据库模型如上图所示:

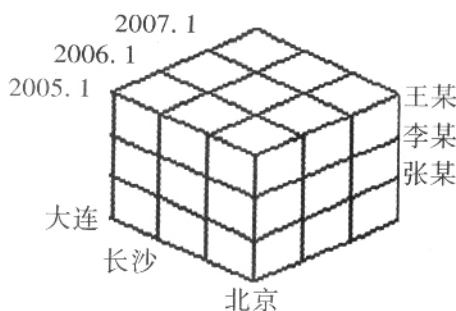


图 4 多维数据模型

在多维数据模型立方体结构中,将大量的数据分成多维结构,作为大数据的子集,对于某一具体应用时,对多维可进行分割,变成多个立方体结构,灵活性强,从而提高了数据分析效率。多维立方体是存储稀疏矩阵的一种行之有效的方法,减少计算量,但是在多维方式的模型中进行查询时,维与事实表之间需进行连接计算,连接时将维表中的维度层次信息转移到事实表中,维表之间并没有关联,因此,对于一些复杂的查询,仍有不足。多维立方体与超立方体相比,终端用户更容易接受超立方体结构,后者可以提供更高水平的报告和多维视图,但多维立方体更适合多维分析人员,因为它具有良好的视图和翻转性及灵活性。

4.2 多维数据模型对 OLAP 分析的支持

OLAP 分析是对多维组织的数据进行数据钻取、聚合、切片、切块、旋转等分析动作,为分析人员提供多种维度、侧面、数据综合度查看数据,从而找出数据之间蕴含的规律,以支持决策。

4.2.1 数据钻取

维度具有层次结构,维度的层次越高,综合度越高,数据细节越少,数据量越少,反之,综合度越低,数据细节越高,数据量越大,如例 4 中,时间维可由年、月、日构成,数据钻取到年、月、日的维度来查看产品销售情况与从年的维度来查看的情况相比,前者的数据要细致的多。

4.2.2 数据聚合

在操作上与数据钻取是相逆的,如例 4 中,对数据综合层次低的以年、月、日为时间维的数据聚合成综合度高的以年为时间维的数据。

4.2.3 数据切块

将完整的数据立方体取其一部分,而得到新的数据立方体,如例 4,取地区维中的(大连、长沙)时间维取(2001、2001)来查看产品销售情况。

4.2.4 数据切片

多维数据由多维度构成,如例 4 中,产品销售维有(大连、北京、长沙),时间维有(2005.1、2006.1),现从维的某子集来查看,如:时间维取 2005.1,地区维取大连、北京来查看。

4.2.5 数据旋转

改变维度的位置关系,为用户提供多角度的观

察,如例 4 中,将时间维和地区维进行交换,形成以时间维纵向,以地区维为横向的报表,除此,在同维度各层次之间也可进行交换。

现有的数据模型能较好的组织数据,但 DSS 的迅速发展需要大量的数据存储及数据间的转换,因此,还不能很好的满足决策系统对信息的需求。

参考文献:

- [1]林杰斌,刘明德. 数据挖掘与 OLAP 理论与实务[M]. 北京:清华大学出版社,2002.
- [2]林宇. 数据仓库原理与实践[M]. 北京:人民邮电出版社,2003.
- [3]彭木银. 数据仓库技术与实现[M]. 北京:电子工业出版社,2002.

Important Data Model in Datawarehouse

PENG Ye - Ping¹, DENG Ting²

1. Department of Information and Engineering, Jishou University, Zhangjiajie, Hunan 427000;

2. Department of Mechanical and Electrical, Vocational and Technical College, Changsha, Hunan 410008)

Abstract: The data model is the foundation of database. Effective data model provide effective and high quality basic data for the OLAP analysis which is based on datawarehouse. Therefore it can support the datawarehouse decision support system in a more effective way.

Key words: OLAP; Snow model; Star model; Multidimensional model

(责任编辑:张荣萍)