数据仓库多维分析模型的设计

李林花　钱越英
（北京应用气象研究所，北京 100029）
E-mail：flower293@sohu.com

摘　要　该文介绍了建立数据仓库多维分析模型的理论基础，提出了一种气象数据仓库多维分析模型的设计方法，并展现了具体的实例，为数据仓库的分析型环境准备了必要的、切实可行的实践基础。

关键词　数据仓库　数据立方体　维表　事实表　星型模型

文章编号 1002-8331-(2004)11-0185-03　**文献标识码** A　**中图分类号** TP311.13

The Design of Multidimensional Analytical Model of Data Warehouse

Li Linhua　Qian Yueying
(Beijing Institute of Applied Meteorology, Beijing 100029)

Abstract: This paper introduces an academic base for building multidimensional analytical model of data warehouse.A designing method of multidimensional analytical model of meteorological data warehouse is presented and illustrated with examples, thus the necessary and feasible practice foundation is prepared for analytical environment of data warehouse.

Keywords: Data warehouse, Data cube, Dimensional table, Fact table, Star schema

1　引言

数据仓库是近年来在信息管理和数据库领域得到迅速发展的一种面向主题的、集成的、随时间变化的、非易失的、用于管理决策支持的数据集合。模型是对现实事物的反映和抽象，它可以帮助人们更加清晰地了解世界，数据仓库模型是数据驻留在数据仓库内的外观蓝图，其设计在业务需求分析之后开始，是数据仓库构建的第一步。

在关系数据库存储数据的基础上，建立决策分析所需的数据仓库模型，需要选择若干对决策活动有重要影响的因素，这些决策分析角度或出发点就构成了数据仓库逻辑结构中的多个维，数据的度量值构成了事实数据的组成部分，从而形成了多维分析模型的两个最基本的数据结构：事实表和维表，它们相结合能够提供给用户一个解决所需业务问题的、清晰、自然的数据视图，以便于分析决策。

多维分析处理分为静态和动态两方面。静态方面包括分析对象和变量的集合，分析的对象被定义为相应变量的函数，每个变量代表空间的一个维。动态方面包括预测、比较、归类、上卷、下钻、数据集成、数据聚合等。维数据从最低的细节级到较高的概要级分为不同的层次，形成了一个家族结构，上一级与它的下一级之间是父与子的关系。在维的不同结构层次中，主要的操作是对数据进行“上卷”和“下钻”。“上卷”是指用户在数据仓库的应用中，从较低层次的数据开始逐步地将数据按照层次进行概括处理，从详细的数据聚集到概要数据的过程。而“下钻”则是指从数据仓库中的高层数据开始逐步向低层数据探索，了解概括性数据的具体细节。

2　数据分析模型方法的形式化描述

一个 n 维数据集合模式是一个三元组 $S(D,T,HR)$，其中：

(1) $D(d_1,d_2,d_3,\ldots,d_n)$ 为维集合，d_i 称为维。

(2) $T(t_1,t_2,t_3,\ldots,t_n)$ 为度量属性集合。

(3) $HR\{(c_1,\in,b_1),(c_2,\in,b_2),(c_3,\in,b_3),\ldots.(c_n,\in,b_n)\}$ 为维层次结构集合，$(c_i,\in,b_i)$ 定义了维的结构层次和聚集约束，c_i 中的每个集合称为维 d_i 的级别，即维成员，$\in$ 表示级别之间的层次关系，b_i 中的每个集合称为维 d_i 的一个维级别的属性。

(4) 多维数据集合是一个实例映射，维的值函数确定了度量属性的值。

设 $DW=\{D_i,T_i,HR|1\leqslant i\leqslant m\}$ 是一个数据仓库模式，DW 的实例是 $dw=\{x_i|1\leqslant i\leqslant m,x_i$ 是 (D_i,T_i,HR) 的实例$\}$。

(5) 一个分析变量为 w，如果存在一个函数（映射）f，定义为 $w=f(d_1,d_2,\ldots,d_n)$，函数域就是 $d_1,d_2,\ldots,d_n$ 构造的多维空间。例如，w 代表气象数据的分析变量，d_1 表示时间，d_2 表示地区，d_3 表示气压，d_4 表示温度，d_5 表示风向，d_6 表示风速等，则 $w=f(d_1,d_2,\ldots,d_6)$ 就唯一地确定了一个具体的天气模式。

(6) 沿着每一维，层次结构被定义。以时间维为例，分为年、月、日、时等级别。

设变量 $c1$ 的值域是 $\{1,2,3,\ldots,24\}$ 代表时；$c2=\{1,2,3,\ldots,24\}$ 代表日；$c3=\{c2_1,c2_2,c2_3,\ldots,c2_{31}\}$ 为月，$c2_1$ 为 1 日，$c2_2$ 为 2 日，依此类推，$c2_{31}$ 为 31 日；$\zeta=\{c3_1,c3_2,\ldots,c3_{12}\}$ 为年，$c3_1$ 为 1 月，$c3_2$ 为 2 月，$\ldots$，$c3_{12}$ 为 12 月。时是数据的最低层次元素，年是层次元素的最聚集部分。如果沿着时间维从时上卷到日，聚集函数为 $w'=f(c2,d2,d3\ldots dn)=\sum_{c1\in c2}f(c1,d2,d3\ldots dn)$，此操作称为上卷，相反的操作称为下钻。

(7) 在多维数据集 $(d_1,d_2\ldots.d_n,w)$ 中，对维 $d_i,\ldots,d_k$ 选定了维成员，即 $(d_1,d_2,\ldots,c_i(d_i$ 维成员$),\ldots,c_k(d_k$ 维成员$),\ldots,d_n,w)$，称为多维数据集 $(d_1,d_2,d_3,\ldots,d_n,w)$ 在维 i~维 k 上的一个切块。当 $i=k$ 时，称为切片，其中，$d_1,d_2\ldots.d_n$ 是维，w 是分析（观察）变量。

作者简介：李林花，硕士研究生，主要研究方向：数据库开发、数据仓库和数据挖掘等。钱越英，高级工程师，研究生导师，主要研究方向：数据库、数据挖掘和软件工程。

3 数据仓库多维分析模型的设计

数据仓库模型设计结构的实现有星型模型、雪花模型和混合模型，最流行的是星型模型。星型结构可以优化数据仓库的查询响应时间，提高查询性能。

星型模型包括事实表和维表。事实表是星型模型结构的核心，由外键和用户希望了解的度量值组成。事实度量值是最终用户在数据仓库应用中所需要查看和分析的细节或聚集数据。维表保存用户查询使用的一个或多个层次关系、成员的类别属性等元数据信息，是相对静态的数据，通过它可以分析数据。每个维表通过一个主键链接到事实表中与此对应的一个外键上。

图 1 是数据仓库的星型模型。

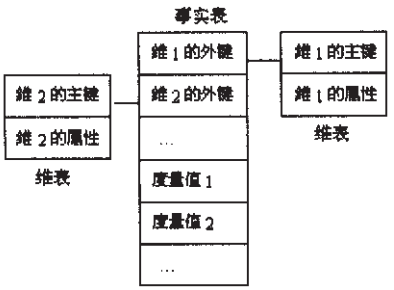

图 1 数据仓库的星型模型

多维数据分析模型把关系模型扩展到立方体模型。数据仓库的多个维构成了数据立方体，维的交点是立方体的顶点，即用户感兴趣数据的度量值。

4 多维分析模型设计实例

气象数据变量繁多，结构复杂，形式多样，包括常规气象资料、卫星云图、雷达、遥感数据和科学数据资料等多种异源数据，格式不统一，具有时空属性，而且观测和接收的数据以每秒上千兆字节的速度海量增长，急需一个巨大的存储空间来存储数据，用高效的分析方法进行信息管理和知识的发现。数据仓库作为异源数据的清理、转换和集成的统一平台，正好适应了这一特点和要求，而多维分析模型的设计是数据仓库模型不可缺少的关键部分。

该文针对气象数据的特点，设计了数据仓库的多维分析模型。

图 2 给出气象数据仓库多维模型星型模型的设计实例。

事实数据对应常规气象要素（气压、气温、湿度、风速、风向、降水量等）的度量值和区站号，维对应地区、时间和各种气象要素。例如，气温维包括了气温、气温的范围与描述等级别层次，其他的湿度、气压、风速、风向、降水量等也有类似的内容。地区维划分为经度、纬度、海拔高度、站名和区站号等，时间维具体到年、月、日、时，是所有数据的时间标志。

按照前述的形式化描述方法，此数据模型可表示如下：

事实表：weather(timeID,regionID,pressureID,temperatureID,humidityID,wind_velocityID,wind_directionID,pressure,temperature,humidity,wind_velocity,wind_direction,precipitation)

维表：dimtime(timeID,year,month,day,hour)

dimregion(regionID,longitude,latitude,altitude,station_name,station_code)

dimpressure(pressureID,pressure,range,description)

dimtemperature(temperatureID,temperature,range,description)

dimhumidity(humidityID,humidity,range,description)

dimwind_velocity(wind_velocityID,velocity,range,description)

dimwind_direction(wind_directionID,direction,range,description)

dimprecipitation(precipitationID,precipitation,range,description)

层次结构：dimtime(hour ∈ day ∈ month ∈ year)，dimtemperature(temperature ∈ range ∈ description)，依次类推。range 和 description 分别是气象要素的抽象概念层次，代表数值的范围和描述。例如气温不是一个具体的数值，而是某个范围（如 10~20℃，20~30℃等），描述为冷、热、适中、暖、温等高级概念层次。其中，timeID，regionID，pressureID，temperatureID，humidityID，wind_directionID，wind_velocityID 是维表的主键。事实表通过每个维的主键值和维表联系在一起。

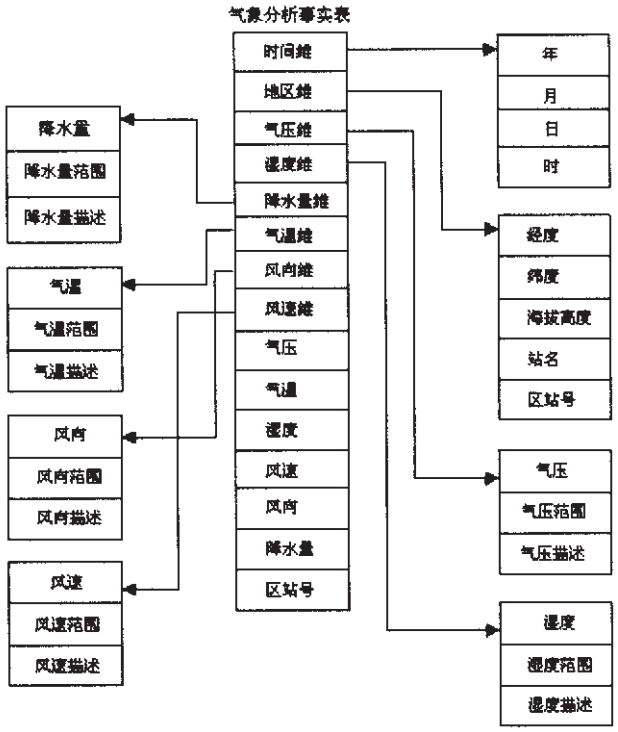

图 2 气象数据仓库多维模型

事实表和维表设计完成后，以此为基础，将星型结构转变为多维立方体结构，其中，星型模型中的事实表对应 OLAP（联机分析处理）中的立方体，维对应 OLAP 中的维。

此设计便于有效地查询和联机分析处理，有利于导出多维空间的一般天气模式。

5 多维分析模型的实现图例

该模型在 Oracle 9i 数据库的基础上，首先采用了 Oracle 9i 的 OWB（Oracle Warehouse Builder）工具进行星型模型的设计，然后在 Oracle 9i 的 OLAP（联机分析处理）中实现了从星型模型向多维立方体的转换，形成了多维分析模型。

图 3、4、5、6 分别显示了星型模型的维表、事实表和多维数据立方体的拓扑结构的图例。

6 结论

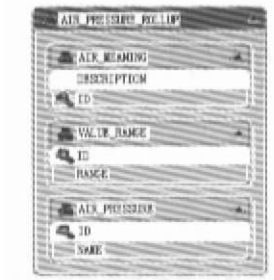

图 3　气象要素维(气压)的设计

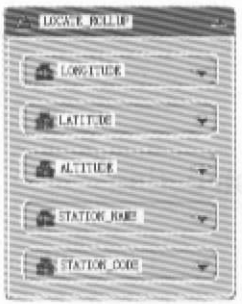

图 4　地理位置维的设计

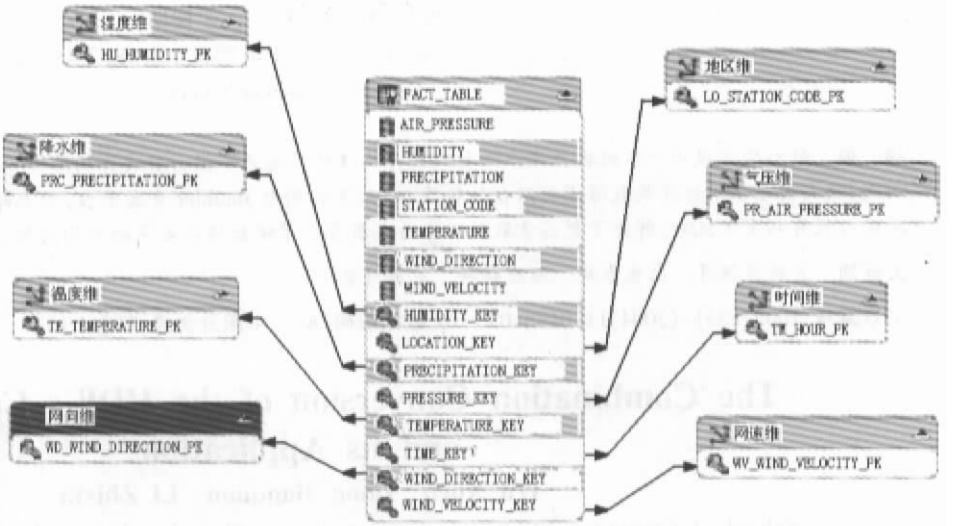

图 5　事实表的设计

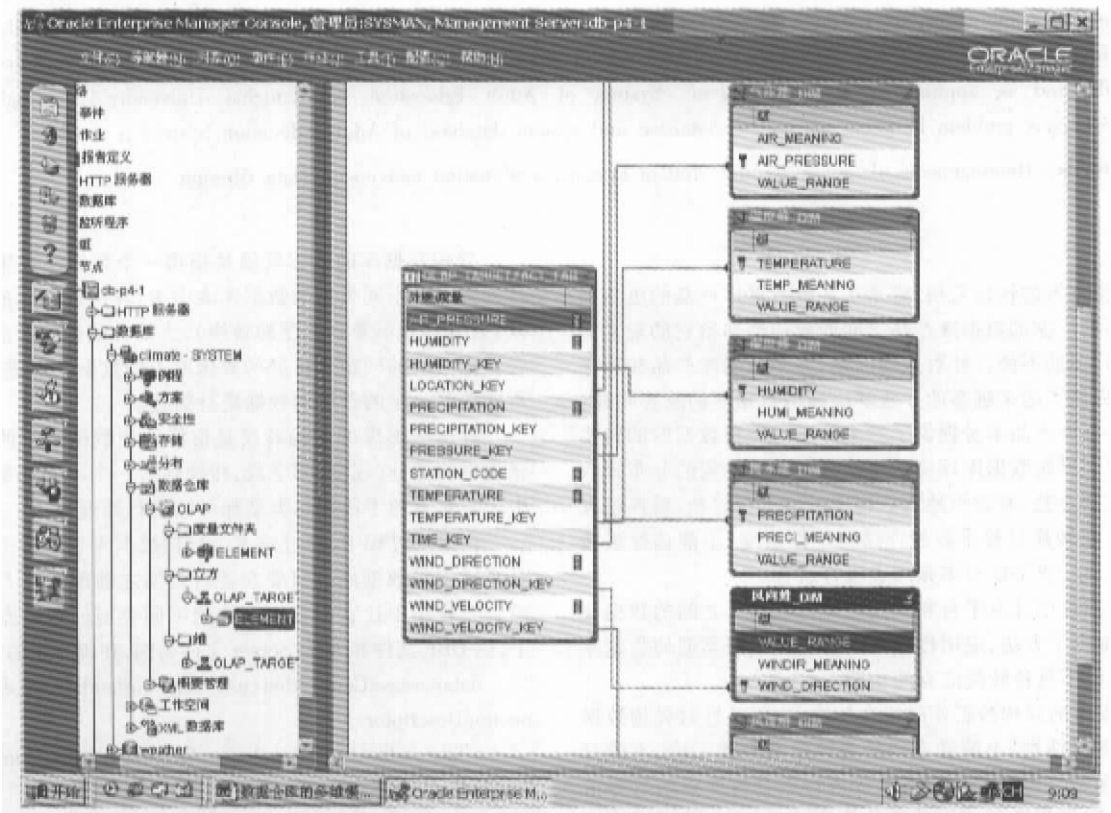

图 6　多维模型数据立方体拓扑结构

多维分析模型非常适合于数据仓库的分析需要,正确而完备的多维分析数据模型也是用户业务需求的体现,是数据仓库项目成功与否最重要的技术因素。该文力图寻找一种适合于复杂的气象数据的多维分析模型设计方法。实验证明,该模型能够满足用户对气象信息多方面的分析应用需求,表达能力强,分析速度快,查询效率高。此项工作是对气象要素分析模型设计的新探索,对数据仓库的构建和后续的数据挖掘研究具有非常重要的意义。(收稿日期:2003 年 6 月)

参考文献

1.王京民.数据仓库与数据挖掘技术[M].北京:电子工业出版社,2003
2.任锦鸾等.数据仓库中数据结构设计方法的研究[J].计算机工程与应用,2001,37(22):116~118
3.Inmon W H.Building the Data Warehouse[M].Wiley,1996
4.Watson H J,Annino D A,Wixom B H.Current Practices in Data Warehousing[J].Information Systems Management,2001,18(1)