学校编码：10384　　分类号＿＿＿＿＿＿密级＿＿＿＿
学号：200228011　　UDC ＿＿＿＿＿＿＿＿＿＿＿＿

厦 门 大 学

硕 士 学 位 论 文

关于数据仓库建模的若干问题研究与实现

Research and Realization on some Problems of Modeling of Data Warehouse

黄 震 华

指导教师姓名：薛 永 生　教授

专 业 名 称：计 算 机 应 用

论文提交日期：2005 年　5　月

论文答辩时间：2005 年　　月

学位授予日期：2005 年　　月

答辩委员会主席：＿＿＿＿＿＿

评　　阅　　人：＿＿＿＿

2005 年 **5** 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文而产生的权利和责任。

声明人（签名）：

年　　月　　日

摘要

目前，人们大都是从如何提高数据仓库上层的OLAP算法和数据挖掘技术来优化连接和聚集查询的效率。然而，一个数据仓库环境的成功与否，很大程度上取决于它的底层模型和标准化程度。基于此，本文以福建省重点科技项目《开放式数据仓库集成环境的研究与实现》为前提，并在原有国内外关于数据仓库模型和规范化资料的基础上进行了这次课题设计。主要针对4个方面的问题进行了研究。

第一，我们基于经典逻辑理论给出了雪花模型形式化定义的 4 个条件，并通过理论和实例分析，对这 4 个条件进行了改进，产生了改进型的雪花模型的形式化定义。并从形式化定义出发给出了机器所能接收的判定算法。这样，可以使机器自动识别一个数据库模式是否是雪花模型，从而减少了人为判断的发错率。

第二，我们给出了从星型或雪花型模式导出对象模型的方法和技术，主要是给出对象模型的类和对象、属性、行为、聚合关系、组合关系和继承关系等的识别标准。

第三，我们以关系数据库的范式理论为基础，结合数据仓库本身的结构和存储特点，给出了减少无效信息和数据冗余的数据仓库范式规则。其中包括 1SSNF、2SSNF、3SSNF 和改进型的 3SSNF，每一个范式都在上一范式的基础上引进了更多的约束，使得冗余和无效信息越为减少。

最后，我们给出了多维数据立方体模型和基于它之上的代数操作，我们证明这些代数操作可以解决目前对以度量属性为查询条件来对维属性的查询问题。同时证明了这些代数操作是最基本的，它们构成的集合满足正确性和最小完备性。

关键词：数据仓库；雪花模型；对象模型；标准范式；立方体

Abstract

Presently, people advance the efficiency of join and aggregation by optimizing the OLAP algorithm and data mining technology. However, it is very important for the success of data warehouse environment to adopt the effective model and the way of standardization. Based on the project of data warehouse, we do the deep research on 4 aspects.

First, we put forward the formalizing definition about snowflake model, which include four conditions, and advance the judging algorithm which can be implemented on computers. In this way, computers can recognize whether a data warehouse model is a snowflake one.

Secondly, we advance the method and technology for producing the object model from a snowflake one. It mostly includes the identifying of class、object、attributes、behavior、aggregate relation、combined relation and successive relation.

Thirdly, based on the normal form theory of relation database and the structure and store characteristic of data warehouse, we bring forward the normal form theory of data warehouse which can reduce the null and redundant data. It includes 1SSNF、2SSNF、3SSNF and ameliorative 3SSNF.

Lastly, we advanced the multidimensional cube model and several algebraic operations. We prove that these algebraic operations can resolve the problem which can query the dimensional attributes from measurement ones. And we prove the set composed of these algebraic operations is satisfied with the minimum and self-contained quality.

Key Words: Data Warehouse; Snowflake Model; Object Model; Normal Form; Data Cube

目录

Index

第一章　绪论

1.1　研究背景

随着数据仓库技术的兴起，越来越多的企业开始建设自己的数据仓库系统，优化企业经营管理，增强企业的竞争力，希望取得较高的回报率。特别是在我国刚刚兴起的制造业信息化的热潮中，将发挥巨大的作用。但是，客观地讲，从数据仓库相关产品的推出到今天，国内企业数据仓库的建设还处在起步的阶段，主要是集中在数据仓库的构建和实施上。由于不能妥善解决数据仓库建设中的一些问题，是不能有效推动数据仓库应用与普及。

数据仓库的显著特点是集成性和随时间变化性，它是通过从不同的操作型数据库和文件源中抽取、转换和加载而得到的，并借助于 OLAP 技术来支持决策。近些年，虽然人们也尝试各种有效的数据仓库模式设计，并提出一些设计方案，但是效果不是很好。主要是因为，目前没有一种标准的方法能够从这些异构数据源中抽取出一种有效模型。其次，数据仓库在组织数据时，还缺乏一种像关系数据库中的范式约束，结果造成了大量无用的信息和数据的冗余。并且在多维数据立方体建模的代数运算方面，也缺乏维属性和度量属性的方便转换，从而造成了从度量属性查找维属性的困难。

由于以上种种的问题和不足，促使我们有必要对这些方面进行比较深入和细致的研究，并且希望能够从中获得可以解决这些问题的方法和技术。

1.2　研究问题的提出

1.2.1　数据仓库多维建模面临的挑战

使用多维数据建模技术面临的挑战有非标准化，使用多重事实表的查询，维表数据量，聚集管理。雪片和数据共享。

① 非标准化：多维数据建模要求数据仓库中的数据通过有效非标准化来避免表的连接。非标准化是一门技术，它经常被用来提高查询性能，但它要求拥有关于数据怎样使用的高级知识。在某些情况中，非标准化技术通过要求 5 倍的数据冗余来达到预期的性能。非标准化易于减少数据仓库的灵活性，

因此限制了在数据仓库上能被执行的查询的种类。

② 使用多重事实表的查询：正如多维数据模型发展为包含多重事实表一样，在一个查询中使用多个事实表是一种趋势。

③ 维表数据量：当与星型结构中的数据库表连接时，首先连接的是维表，然后这些表的交互根据这个事实表进行。如果维表比较小，或者查询中维表满足要求的数据行的数目有很大的选择性，会根据这个事实表产生一个相对小的中间结果集。在另一方面，一个或多个大的维表会产生一个巨大的中间结果集，这显著增加了查询响应时间。

④ 聚集管理：聚集是根据已知查询预先计算好的结果。如果生成太多的聚集表，会遭遇到聚集表爆炸问题。在一个只包含一个事实表和 10 个维度属性的数据模型中，可能建造超过 3 百万个聚集表。只有极少的组织有这个处理能力来管理所有来自这个小模型的潜在聚集。

⑤ 雪片：使用层次结构意味着它能显著地增长，例如许多地理级别能被用来对商店进行分组。如果使用雪片技术，并在每一层上增加维表，这可能使得查询性能进一步恶化。在这种情况下，可以运用非标准化技术，把地理维表统一在商店维表中，从而减少所有地理层次表的可能连接。

⑥ 数据共享：为了优化性能，一些多维数据模型在实现时把每个星型结构分割成独立的数据库。因为某些维表不只在一个星型结构中使用，所有必须在每个数据库中复制这些表。这项技术提出了挑战，必须为这些数据的额外备份提供附加的存储空间，以及保持这些备份表的同步。

1.2.2 数据仓库规范化问题

在数据仓库刚起步的阶段，由于种种原因，很多开发人员发现在数据仓库建模时，不能用关系数据库的规范化理论来约束数据仓库模型。因为为了提高在数据仓库之上的 OLAP 操作和决策分析的效率，必须使得数据仓库中的数据要有冗余，并且要非规范性。但是近年来，很多专家学者发现数据仓库设计绝对是一个适合于使用规范方法的领域。关于为什么规范化可以产生数据仓库的最优设计有几个很好的原因：

a. 规范化方法可以带来灵活性。

b. 规范化方法很好地适应于粒度化的数据。

c. 规范化方法不是对任何给定处理需要集合都是最优的。

d. 规范化方法能很好地与数据模型相匹配。

目前由于数据仓库的规范化技术还不是很成熟，所有到目前为止，还没有提出一套象关系数据库那样比较完整的范式结构。本课题的一个重要内容就是研究数据仓库的范式，并设计出一套比较完整的范式结构。同关系数据库一样，我们也要给出数据仓库的第一范式，第二范式，第三范式等范式规则。这些范式在一定程度上能够规范化数据仓库，使数据冗余达到比较小的状态，为 OLAP 操作和决策支持分析提高良好的且较规范的数据仓库模型，从而提高 OLAP 操作和决策支持分析的效率。

1.2.3　数据仓库模式的对象模型

不管数据仓库的模式如何，规范化程度多大，它最终都是为数据仓库的软件开发服务的。由于在软件的分析和设计时，都要采取某一种形式的建模方法。目前，有三种不同的建模方法：面向数据流、面向数据结构和面向对象的方法。因此，怎样把数据仓库模式转换为软件分析模型是对整个数据仓库开发周期至关重要的。由于面向对象建模方法越来越流行，并且数据仓库模型的主题和面向对象中的类和对象有着很大的相似性，所以我们研究如何把数据仓库模式导出转换成对象模型。

1.2.4　数据仓库多维数据立方体模型和代数运算问题

先前关于数据仓库多维数据立方体模型和代数运算的研究，主要集中在把维表和事实表所处的位置进行显式的分离和固定，在这种处理方式，只能通过维表的属性来聚集查找事实表中的度量属性；而很难通过事实表中的度量属性组来标识出维表中的属性集，这种代数方式对于构造SQL语言时存在这从维表到事实表的单方面的查找和聚集行为，而摒弃了从事实表到维表的这种逆操作。

其次，以前的关于代数运算的研究，都没考虑基本操作，即能够构成最

小正确完备性的代数运算，而是列出了很多种代数运算，而里面的很大部分的代数运算可以有若干个基本操作组合而成，在这种情况下，就造成了代数运算的冗余和复杂性。

1.3 本文所做的工作

对国内外多维数据仓库建模方法及相关资料的认真研究和分析，特别对国外那些比较有影响的数据仓库模型的研究，从中找出各个模型的优点和缺点。基于此，本文所要做的工作包括：

① 由于现在数据仓库借鉴成熟技术的关系数据库的技术，所有现在大都采用基于 ROLAP 模式的星型模型和雪花模型，其中星型模型是雪花模型的特例，雪花模型是星型模型的拓展。到目前为止，每个人在思维上有这星型模型和雪花型模型的概念，能够判断什么是星型模型，什么是雪花型模型，但是在模型描述方面还不能实现，特别不能用数学和逻辑定义给出一个准确的结论，所以本次研究的一个重点是给出星型模型和雪花模型的准确定义，有一个正规的判断标准。

② 目前由于数据仓库的规范化技术还不是很成熟，所有到目前为止，还没有提出一套象关系数据库那样比较完整的范式结构。本课题的一个重要内容就是研究数据仓库的范式，并设计出一套比较完整的范式结构。同关系数据库一样，我们也要给出数据仓库的第一范式，第二范式，第三范式等范式规则。这些范式在一定程度上能够规范化数据仓库，使数据冗余达到比较小的状态，为 OLAP 操作和决策支持分析提高良好的且较规范的数据仓库模型，从而提高 OLAP 操作和决策支持分析的效率。

③ 近几年，人们在数据仓库的代数操作研究方面作了很大工作，并逐步在完善涉及数据仓库操作的代数定义。但是所有的研究工作都按照着某一相同的模式进行：以某些维属性为条件，来显示某些度量属性。这样的话，数据仓库的应用就受到了某些限制，比如我们就不能以度量属性为条件，来显示某些维属性，基于此这次课题的研究的一个重点，就是突破常规的限制，