

数据仓库的多维数据管理与分析技术的应用研究^{*}

李国珍¹, 郭清顺²

(1. 石油大学, 北京 102200;

2. 中山大学网络与教育技术中心, 广东 广州 510275)

摘 要: 数据仓库技术是进行大规模数据管理和分析的有效工具, 但是如何建立一个有效的数据仓库和数据分析系统仍然存在许多问题。针对数据仓库建设的元数据获取、多维数据管理等数据仓库设计技术与数据分析技术做了深入探讨, 给出了多维数据设计的评价标准与方法, 通过在大庆油田数据仓库建设实践证明其方法是十分有效的。

关键词: 数据仓库; 多维数据分析; 油田开发

中图分类号: TP311.13 **文献标识码:** A **文章编号:** 0529-6579 (2002) S1-0045-04

数据仓库和数据挖掘技术是 90 年代初期出现的新的数据管理和数据分析技术, 不仅被理论界和学术界视为是对数据库技术和人工智能技术的重大发展, 而且也被企业界看成为是能够为其带来巨大效益的应用领域。把数据仓库技术和数据挖掘技术有机结合, 利用数据仓库管理和整合数据, 用数据挖掘技术挖掘数据之间的内在关系, 对于企业进行智能化管理是十分重要的, 中国加入 WTO 后增加了对中国企业的挑战, 采用数据仓库管理与数据分析技术无疑会为我国的工业企业带来巨大的经济效益。数据仓库技术是高层数据分析技术, 是信息管理与分析的前沿技术, 有许多问题尚待研究和解决, 目前它是国际上的一个研究热点, 同时也是一个非常有前景的研究领域。数据仓库支持数据分析的基础是多维数据库设计与元数据的管理与获取, 所以本文以此为题与读者共同探讨。

1 数据仓库的概念属性

数据仓库 DW (data warehouse) 概念的形成是以美国工程信息学家、Prism Solutions 公司副总裁 W. H. Inmon 在其 1992 年出版的著作《建立数据仓库》(Building the Data Warehouse) 一书提出而作为标志的。在该书中, 他对数据仓库的定义为: “数据仓库是支持管理决策过程的, 面向主题的、集成的、随时间而变的、持久的数据集

合”。这一概念和在 DW 之前引入的联机分析处理 OLAP (on line analysis processing) 方法较好地解决了在信息技术发展中存在拥有大量数据及如何利用其中有价值信息, 并解决当今信息社会中存在的 “Data rich—Information poor” 的严重问题。

数据仓库的基本概念属性就是把有关主题的历史数据和当前数据 (包括内部和外部数据, 结构化和非结构化的数据) 收集到一个中央仓库中以便进行分析处理。它是支持管理决策过程的、面向主题的、集成的、随时间而变的、持久的数据集。数据仓库技术是当今信息管理技术的主流, 是促进企业正确管理与决策的重要决策支持工具。

2 获取并管理元数据

整个数据仓库的数据组织结构是由元数据来组织的, 元数据是 “关于数据的数据”, 它不包含任何业务数据库中的实际数据信息。如在传统数据库中的数据字典就是一种元数据。但在数据仓库中, 元数据的内容要比数据库中的数据字典更丰富、更复杂。元数据目录是数据仓库的 “路径图”。元数据至少应包含如下一些信息: 数据结构、数据综合算法、从基于 OLTP 的业务环境到数据仓库环境的规划等。设计一个描述能力强、内容完善的元数据, 是建立有效数据仓库的

^{*} 收稿日期: 2002-03-21

作者简介: 李国珍 (1976—), 女, 博士研究生; E-mail: yaoyuedong@163.com

重要环节。由于元数据在数据仓库中扮演了重要角色,因此,元数据的设计在整个数据仓库的设计中占有重要地位,并作为数据仓库设计的一个重要组成部分。

3 多维数据库设计方法与标准研究

传统数据库存在的缺陷是其分析软件是静态的,不适于进行动态数据分析;此外,尽管关系数据库提供了根据不同的视图或需要进行数据选择、比较及综合的能力,但数据可在多个维上以不同的格式、不同的定义出现的这一事实,充分暴露了传统关系数据库系统在数据分析能力上的局限性。缺乏多维分析能力,这正是关系数据库系统的缺陷。为实现数据仓库和 OLAP, E. F. Codd 在有关数据仓库与 OLAP 的概念、特征及其产品评价的基本准则中,明确提出数据仓库和 OLAP 模型必须提供多维概念视图。这是因为从用户分析员的角度看,整个企业的视图本质上是多维的。而且,随着企业决策分析目的的不同,决定了分析和衡量企业的数据总是从不同的角度来进行的,所以企业数据空间本身就是多维的。因此,数据仓库和 OLAP 的概念模型也应是多维的。在作为数据分析工具的数据仓库和 OLAP 中,应如何提供直观的多维分析模型进行分析、设计和维内到维间的计算是必须加以充分考虑的。这种多维模型可以使最终分析比传统的关系数据库系统采用单一维模型更简单、直观的方式操作多维数据。例如,我们通过对多维分析模型进行切片、切块、旋转和向上与向下挖掘等操作就可以轻松地完成传统数据库操作方法必须用大量时间和很大系统开销代价才能完成的工作。

多维数据分析是指对多维形式组织起来的数据采取切片、切块、旋转等各种分析操作,以求剖析数据,使最终用户能从多角度、多侧面地观察数据库中的数据,从而深入地了解包含在数据库中的信息、内涵。多维数据分析方式十分适合人的思维模式,它可以减少混淆,降低数据造成错误的可能性。维数确定的评价标准主要有:①符合人们进行数据分析的思维习惯;②有较好的易理解性;③便于组织和管理元数据;④便于数据更新或重组;⑤便于用户使用;⑥能支持多重粒度,便于数据的“概化”和“细化”;⑦支持数据的切片、切块和轴(维)

旋转;⑧能够表述结构化和非结构化数据;⑨有较好的(维)的扩充性;⑩要能够真正表征主题对象的内在属性(类);⑪要能够真正刻画主题对象的外在特征(类);⑫每一维都要有对应的元数据;⑬每一维都要有实际的应用或分析的意义。

上述 13 条准则是作者在课题的研究过程中提出的,按照这 13 条准则,我们确定了大庆油田开发数据仓库的“维”,有关详细设计见本文的第 4 节。

4 油田开发数据仓库的概念模型

油田开发数据仓库的主要功能在于为油田开发指标分析、油田剩余油分布和潜力预测提供数据支持,因此数据的需求在于准确把握油田每口井、每个区块的产油、产水状况,分析产油产水的变化趋势、各项开发指标,从而为今后的油田开发进行合理有效的部署。所需数据应包括:①油水井的固有信息(静态信息);②油水井的小层数据(静态信息);③油井产油量、产水量、压力等信息(动态信息);④水井的注水量、压力;⑤油水井按照不同时间综合度导出的综合信息;⑥油田区块单元的划分信息;⑦按照不同区块、不同时间导出的综合数据。

针对以上的数据种类,将数据仓库的数据边界定为包含原有的单井基础信息、油井月数据、水井月数据、单井小层数据、钻井地质信息等原始数据以及在此基础上导出的综合指标数据的集合。

根据对原有数据库中数据表的结构以及数据状况的分析,并考虑到油田开发数据的查询统计、分析预测及数据采掘的要求,在系统边界的划分基础上,数据仓库中的一切数据将针对油田开发指标分析、油田剩余油分布和潜力预测两个主题,围绕油水井的各种生产信息来进行组织。包括按油田、区块等不同地域范围进行的油水井生产指标等的统计。油田、区块、井之间通过井号相互关联。

5 油田开发数据仓库的逻辑模型

数据仓库的逻辑结构设计时考虑基础性信息和与时间有关的统计性信息两部分。

(1) 基础性信息:包括单井信息、单井小层数据、油水井井史、措施效果数据、动态监测数

据、区块信息等，其逻辑结构采用股份公司的开发数据库逻辑结构。

(2) 统计性信息：以时间、区域划分、行政单位划分、油层划分作为不同的维，分类统计相应的指标数据。

大庆油田包含有 40 000 多口油水井，每个月有近 40 000 条月井史记录进入数据库；油田经过 40 年的开发，其数据量是非常大的。考虑到数据仓库所具有的大数据量以及在进行指标分析预测时的数据查询类型，粒度划分定义的综合层次要丰富一些。按时间划分的起始点是月数据，然后包括年统计数据；同时将油田区块单元的划分考虑进去，形成在不同的时间段上的按区块单元的统计数据。

5.1 维的确定

- 时间维：月、年 2 个粒度。
- 区域维：油田、区块、小区块、井号 4 个粒度。
- 行政单位维：厂、矿、队 3 个粒度。
- 油层维：油层组、小层号、细分层号 3 个粒度。

5.2 层次的划分

每一个维都具有一个或多个层次，其划分的方法是采用包含关系，即按照粒度范围的从小到大进行排列。

5.3 维表与事实表

对于关系模式的定义要按照当前实施的主题进行模式划分，形成多个具体的维表和事实表，并确定各个表的关系模式。目前，油田开发数据

仓库已经建立了 4 个维表和 8 个事实表。

按照油田开发数据仓库逻辑模式的定义，经过数据检查、数据转换、数据加载，油田开发数据仓库已经建立了 4 个维表和 8 个事实表。数据仓库的事实表与维表之间的逻辑关系可以用下面的星型图表示（图 1，以 f_sjsc 表为例）。

6 油田开发数据的多维分析

根据数据旋转理论和油田开发数据分析的特征，基于数据仓库中维的数据综合查询、分析，可以充分利用多维数据分析技术对油田开发数据仓库中的数据进行统计、汇总以及分析，为用户提供了一个有效的数据分析工具。

我们利用 Oracle 提供的数据库及分析工具 Oracle Discoverer，实现了针对油田开发数据按照时间、区域、单位等不同的数据维进行多维数据查询、统计、分析。所建立的查询、分析模式具有独特的数据查询预测、自动汇总和重新导向等特点，在提供数据统计报告的同时能够将所查询统计的数据以表格或图形方式显示。某区块油田开发的月产油量、月产水量的分析查询统计实例如图 2 所示。油田开发数据仓库采用多维表的设计方式，以油田开发信息为基础，既支持传统的数据查询及报表等功能，又支持多维数据分析、数据挖掘功能，从而为油田各级管理及技术人员提供生产动态、决策分析所需要的有价值的信息。其维表的设计由常见的统计分类方式确定，事实表包括了油田生产的基础信息及静态地质信息。

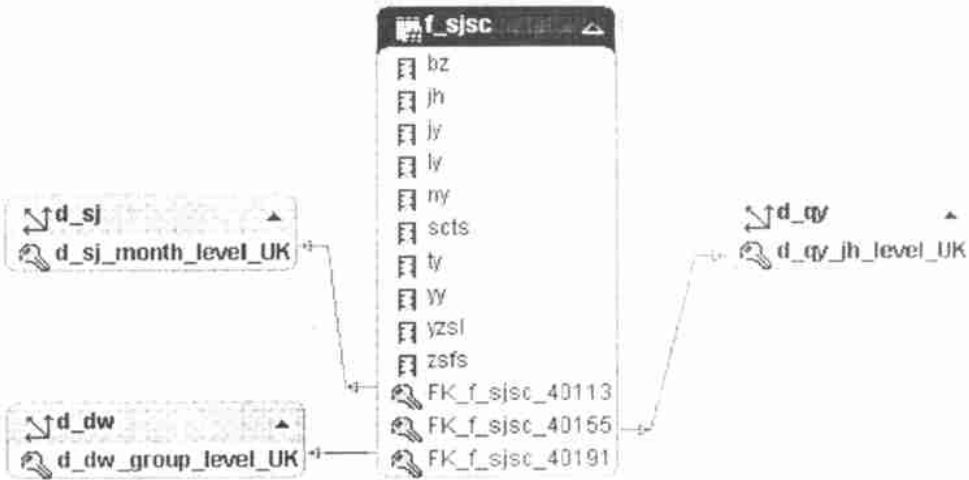


图 1 逻辑模型

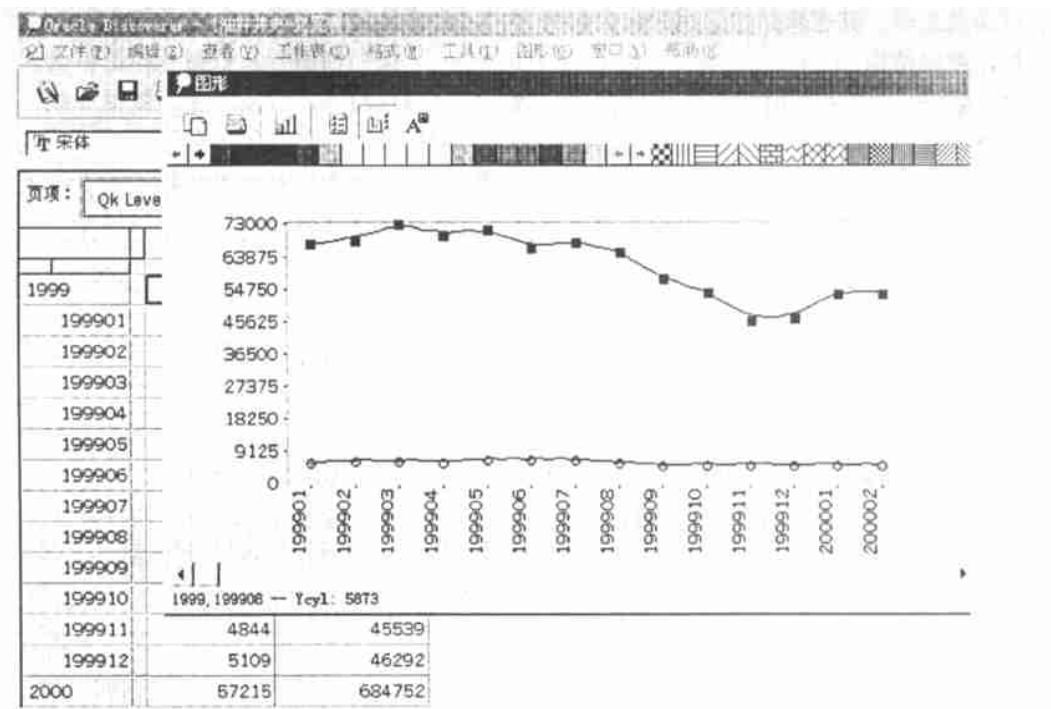


图 2 数据统计表及图形

7 应用效果与结论

在 2000 年末对油田开发生产情况进行分析统计的过程中，分别按照井号、时间、区块等不同的分类统计方式统计了数据仓库中各区块的投产井数、产油、产水、产液、综合含水、注水、压力等一系列开发指标，并在此统计数据的基础上进行了油田注水开发效果评价、油田开发指标预测、油田开发综合调整措施优化和剩余油潜力预测等各类多维数据分析。在对数据的查询、提取、统计和进行有效的多维分析过程中，与往年

采用二维关系数据表相比，其所需时间减少一半以上，解决了数据难以快速统计、汇总、分析的问题。

参考文献：

[1] 马玉书. 人工智能及其应用[M] . 北京: 石油工业出版社, 1998.

[2] 范明, 孟小峰. 数据挖掘——概念与技术[M] . 北京: 机械工业出版社, 2001.

[3] 王志海. 数据仓库[M] . 北京: 机械工业出版社, 2001.