石化市场数据仓库建设初探

姜洪安　（中国石化总公司信息中心　100029）
黄涛　成理宙　（中国科学院软件所对象中心　100080）
王　葵　（北京外企服务总公司　100027）

摘要：本文基于石化企业发展的需要，在比较传统数据库与数据仓库各自技术特点的基础上，着重阐述了石化市场数据仓库的实质、特征、基本体系结构及建设步骤，提出了建设过程中应注意的几个问题。

关键词：数据仓库　石油化工

一、石化市场数据仓库的实质与特征

随着传统的数据库技术应用的日益发展，人们开始尝试对数据库中的原始数据进行再加工，形成一个综合的、面向分析的环境以支持科学决策的产生。由此，数据仓库的思想、技术和软件产品逐步形成。

数据仓库的提出是以传统的关系数据库、并行处理和分布式技术的飞速发展为基础，其目标是解决在信息技术发展中存在的拥有大量数据资源，但有用信息贫乏（Data Rich - Information Poor）的问题。数据仓库概念的创始人 W. H. Inmon 给数据仓库作出的定义是："数据仓库就是面向主题的、集成的、稳定的、不同时间的数据集合，用以支持经营管理中的决策制定过程"。从广义上说，数据仓库就是"来自一个或多个数据库的数据的拷贝"。由此，我们可以把石化市场数据仓库定义为：石化市场数据仓库是石化企业内部、国内石化市场、国际石化市场、石化产品进出口等专用性数据库系统的运作数据和事务数据的中央仓库，这些数据按统一的抽取方法抽取出来，并经过了归化、平衡、协调和编辑。

1. 石化市场数据仓库的实质

首先，应当明确，数据仓库是传统数据库技术的一种新的发展和应用，其实质仍是计算机存储数据的系统，只不过它存储的数据在量上和质上都与专用性数据库有所不同。

其次，数据仓库中存放的数据是分析用的数据，它不存放与分析无关的纯操作性数据。

第三，数据仓库存储的大量数据（包括历史数据、当前数据和综合数据等）随着分析主题的增加而增加，因此数据仓库不是一成不变的，而是处于发展变化之中。当然，数据仓库中的数据并非只增加不减少，对分析主题不再有用的数据、一些经过综合后遗弃的细节数据都应当被清理掉。

第四，石化市场数据仓库不是对现有专用性数据库系统的替代。数据仓库侧重于综合分析，专用性数据库侧重于一般性数据处理。专用性数据库是数据仓库的基础，没有专用性数据库就没有数据仓库。

第五，来自专用性数据库的数据在进入数据仓库前要进行转换和校验，因而数据仓库中的数据具有一致性的特点。同时，由于数据仓库是为信息分析提供支持的，因此它对信息分析人员而言是只读数据库，即信息分析人员通过前端工具和应用程序来访问数据仓库，但不能对其中的数据进行任何修改，只能定期刷新。

2. 石化市场数据仓库的特征

（1）面向主题性。石化市场数据仓库应是面向石化产品、原材料市场这一大的主题。主题是一个在较高层次将数据归类的标准，它是与传统数据库面向应用相对应的。石化市场主题又由多个分析主题组成，每个分析主题基本对应一个综合分析领域，通过一组有关联的表来实现。例如，石化总公司要统一协调某种石化产品的保底价，保底价定多少合适，就是一个分析主题。要制定出合理的保底价，必须了解该种产品在国内的生产企业情况、生产量情况、产品质量情况、生产成本情况、出厂价情况等，这些数据来源于企业的专用性数据库系统。我们还应了解该种产品的主要销售市场情况、主要客户情况、市场价格情况等，这些数据来源于市场及客户数据库系统。此外，我们还需要掌握该种产品的进口量、进口价格、经营单位、贸易方式、国际市场价格、主要供应商等，这些数据来源于海关进出口数据库系统和国际石化市场数据库系统等。

（2）数据集成性。石化市场数据仓库的第二个重要特征在于其数据的集成性。前面我们已经提到，数据仓

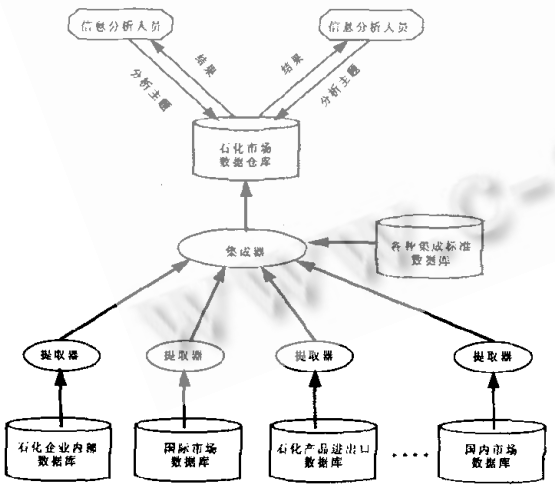
库的数据主要作分析用，分析用数据的最大特点在于它不局限于某个具体的操作数据，而是对细节数据的归纳和整理。例如，我们要对某种石化产品的市场占有率进行分析，必然要用到进出口数据库系统，但我们不是把这种产品的每一个细节数据（如经营单位、贸易方式等）都搬到数据仓库中来，可能只是取其某一年（或几年）的进口量这一经过统计以后的数据，这些数据就变成了集成数据。数据集成是数据仓库建设过程中最关键、最复杂的一步。我们不仅要统一原始数据中的所有矛盾之处，如同名异义、异名同义、单位不统一等，而且要将这些数据统一到数据仓库的数据模式上来，还要监视数据源的数据变化，以便更新和扩充数据仓库。

(3)数据稳定性。数据仓库主要为信息分析提供经过综合的、集成的、面向某一分析主题的数据，这些数据原则上不允许信息分析人员直接对数据执行修改或删除操作，由此也保证了石化市场数据仓库的稳定性。

(4)较强时间性。因为数据仓库中的数据常常要被用作趋势分析，必须有足够多的时间里的数据才能进行，因此数据仓库中的数据一般存放时间较长，甚至达数十年。这就要求这些数据必须建立含有时间项的码键，以便标明数据的历史时期。

二、数据仓库的体系结构与建设方法

1. 石化市场数据仓库的基本体系结构



其中：

(1)提取器主要负责转换数据模式，监视专用性数据库上的数据变化；

(2)集成器主要负责按数据仓库的各种规则（如一致的命名转换、一致的变量度量、一致的编码结构、一致的数据物理属性等）将信息加载到数据仓库中，其中可能要进行过滤、汇总，或与其他信息源的信息合并，然后把新信息正确地集成到数据仓库中。

2. 石化市场数据仓库的建设方法

数据仓库的建设不是一蹴而就的，它是一项复杂而艰巨的工作，目前尚没有一种标准的建设方法，因此建设数据仓库应根据企业的具体情况而定。石化市场数据仓库的建设包括以下几个方面。

(1)组织数据源。从上面定义的石化市场数据仓库的基本体系结构中可以看出，数据仓库的数据不仅来自信息中心目前已经建立的专用性数据库，而且更多的是来自石化总公司其他业务管理部门和所属企事业单位已经建立的或正在建立的专用性数据库系统。要把这些数据组织起来并非易事，原因在于一是人们对信息共享性的认识程度不够，不愿意将数据提供出来；二是要整理归类出纯操作性数据和分析用数据，包括数据粒度的划分等；三是如何从信息源获取信息，并将它们组织，集成到数据仓库中，主要包括：如何按数据仓库主题从数据源中提取相关的数据，将它们转化为数据仓库模式；如何解决来自不同信息源的数据的重复和不一致性等问题；如何建立和完善元数据字典。属于技术方面的问题可以通过软件来解决，属于其他方面的问题则需要采取管理手段、行政措施以保证数据源。

(2)建立数据模型。通过数据模型，我们可以得到整个石化市场完整而清晰的描述信息。数据模型应面向分析主题建立，同时又为多个专用性数据库系统的数据集成提供统一标准。石化市场数据仓库的数据模型应包括：石化市场的分析主题域、各主题之间的联系、描述主题的码键和属性组等。

(3)设计数据仓库结构，确定数据仓库应用环境。以数据模型为基础，剔除纯操作性数据，增加部分描述属性（如时间属性等），定义各种表结构及表间的联系。同时，应确定数据仓库的应用环境，包括开发环境和运行环境，如软硬件平台、软件工具、网络环境等。

(4)制定统一标准。石化市场数据仓库中的数据来源于多个已有的专用性数据库系统。这些数据库系统基本是面向应用建立的，不能完整地描述市场分析主题，也容易产生数据之间的不一致（如命名、结构、单位不一致等），甚至出现同名异义、异名同义等情况。所以，必须根

据分析主题的需要及现有专用性数据库系统的实际情况制定出一套标准来。这套标准的特点是：数据完整、准确、及时并适合于数据库到数据仓库的转换。

(5)引进/开发软件工具。主要有四类软件工具需引进或开发，它们是：数据抽取类工具，数据管理类工具，数据挖掘类工具，数据分析类工具。

(6)建立第一个分析主题，在反馈和循环中逐渐建立其他分析主题。建设数据仓库是一个艰巨而漫长的复杂过程，也是一个逐步建设、逐步完善的过程，因此可以考虑先以一部分数据生成第一个分析主题，以便设计人员能够容易且迅速地对已做工作进行调整。这样做的好处在于一方面使数据库尽快投用、尽快见效，为建立后续的分析主题树立样板，另一方面又能通过信息分析人员的使用发现新的问题，提出新的需求，便于继续对系统进行改进、扩展，并将更多的分析主题加入到数据仓库中。

三、建设市场数据仓库应注意的问题

1.提高认识，放眼未来

有人曾说过这样一句话：答案标志着旧问题的解决和新问题的开始。从事石化市场信息分析的业务人员，为了能够准确分析出“是什么”和“为什么”，要经过多次反复的查询、制表、分析。因此，我们不能把石化市场数据仓库看作是一个静止不变的产品，而应当把它看作是一个动态的、不停的变化过程和解决方案。可以说，石化市场数据仓库从实施的那一天开始，就没有终结的时候。它的建设过程，正是石化市场信息分析水平从简单到复杂、从局部到全局的提高过程。我们必须从长远的角度建设数据仓库，制定出切实可行的实施计划，必须保证系统是灵活的、可扩充的、模块化的，以便有足够的能力适应系统的不断增长。在这里需要强调的一点是：数据仓库的建设投资很大，是一项长期、复杂并面临很大风险的工作，必须谨慎对待。

2.分析工具的重要性

建立数据仓库以后，如何使用它，即使用、管理数据仓库的问题。主要是使用仓库进行决策支持，或采用数据挖掘技术发现各种数据形式的内在联系和人们所未知的知识。

软件工具是整个石化市场数据仓库发挥作用的关键，而数据分析类工具则是关键中的关键。在目前还缺乏既是数据库专家又是石化市场信息分析专家的复合型人才的情况下，要满足信息分析人员多变、复杂的需求，使他们能够很容易地查询到所需要的各种数据，并能直观的、方便的制表，作分析图形，成为石化市场数据仓库能否建设成功的关键环节。我们上面提到的四类软件工具中，其他三类基本是由数据仓库设计人员或维护/管理人员进行操作，而只有数据分析类软件工具是直接面向信息分析人员的，因而也被称之为前端工具。前端工具是石化市场数据仓库的重要组成部分。对于信息分析人员而言，如果仅拥有石化市场数据仓库，而没有高效的信息分析工具，就如同守着一座储量丰富的金矿而不知如何采掘。分析型工具大体分两种模式，即验证型和发掘型。验证型工具的主要作用是从数据仓库中发现事实，以验证或否定用户的假设。它又分为可视化工具和多维分析工具。发掘型工具主要负责从大量数据中发现数据模式、预测趋势。这类工具以数据挖掘软件为代表。由于数据分析类软件工具的前端直观性，决定了前端工具必须给用户提供熟悉的业务术语，使之不必成为数据库专家也能对数据仓库运用自如，获得关键性信息。

3.项目组的构成

要保证石化市场数据仓库的建设成功，必须有一个强有力的、高效的项目组。该项目组应由三方面的人员组成。一是石化市场信息分析人员，他们主要负责确定分析主题及需要的数据，使用前端工具进行查询分析，并对数据仓库提出改进意见；二是数据仓库设计人员，他们主要负责数据仓库的体系结构、物理存储结构及逻辑结构的设计，理解信息分析人员提出的分析主题的含义，定义语义层；三是数据仓库管理人员和维护人员，他们主要负责源数据的提取、转换，并把这些数据加载到数据仓库中，他们还负责记录数据仓库整个建设过程的所有文档的编辑、整理、管理工作。

参考文献

[1] 王珊，罗立，“从数据库到数据仓库”，“计算机世界”第 101 期；

[2] 王珊，刘方，“创建数据仓库的方法、模型与步骤”，“计算机世界”第 101 期；

[3] Microsoft 著，陈河南，贺军译，“客户－服务器系统分析与设计”；

(来稿时间：1997 年 12 月)