文章编号: 1001-0920(2004)12-1378-05

企业级数据仓库设计方法及其实施的关键因素研究

鲍玉斌[1], 史 捷[2], 王大玲[1], 嵇 晓[1], 于 戈[1]
(1. 东北大学 信息科学与工程学院, 辽宁 沈阳 110004; 2. 中铁九局集团有限公司 信息中心, 辽宁 沈阳 110013)

摘 要: 数据仓库是支持企业全局决策的有效技术. 首先讨论了数据仓库的基本结构和组成, 并给出了数据仓库系统的形式化描述; 然后给出了基于软件工程思想的具有 7 个阶段的数据仓库设计方法, 并详细讨论了各个阶段的主要工作和技术; 最后, 给出了成功建设数据仓库的关键因素的分类.

关键词: 数据仓库; 体系结构; 设计方法; 关键因素
中图分类号: TP311 **文献标识码**: A

On the design of enterprise-wide data warehouses and the critical factors in the development

BAO Yu-bin[1], *SHI Jie*[2], *WANG Da-ling*[1], *JI Xiao*[1], *YU Ge*[1]
(1. School of Information Science and Engineering, Northeastern University, Shenyang 110004, China; 2. Information Center, China Railway No. 9 Group Co. Ltd, Shenyang 110013, China. Correspondent: BAO Yu-bin, E-mail: baoyb@mail.neu.edu.cn)

Abstract: Data warehousing is an effective technique for decision-making support in range of whole enterprise. The architecture and the component of data warehouse system are discussed, and its formal description is presented. Then, the framework including seven phases for designing data warehouse system is presented based on software engineering. And the main tasks and techniques of each phase are discussed. Finally, the classification of the critical success factors for data warehouse design is proposed.

Key words: data warehouse; architecture; methodology; critical factors

1 引 言

数据仓库是面向主题的、集成的、随时间改变的、持久的数据集合, 主要用于支持经营管理中的决策制定过程[1]. 数据仓库系统是企业有效利用庞大信息资源的一个很好的解决方案. 这不仅是因为数据仓库能提供庞大信息资源的有效管理, 而且更为关键的是数据仓库技术一改以往数据库技术的"以数据为中心"的理念, 它强调"以信息、业务为中心, 以决策为目的". 数据仓库是企业开发各种应用系统, 如企业资源规划(ERP), 客户关系管理(CRM), 供应链管理(SCM), 数据挖掘(DM), 联机分析处理(OLAP)等的基础. 因此, 构建企业级数据仓库是企业成功实施其他系统的关键.

文献[1~3]中提出了数据仓库设计方法, 讨论了数据仓库的最核心部分的设计, 但没有从整个数据仓库工程的角度出发. 因此, 本文在给出数据仓库形式化描述的基础上, 通过某大型企业数据仓库的开发实践, 提出企业数据仓库系统设计方法, 总结了成功开发数据仓库系统的几个关键因素.

收稿日期: 2003-08-04; **修回日期**: 2004-06-15.
基金项目: 国家自然科学基金资助项目(60173051).
作者简介: 鲍玉斌(1968—), 男, 吉林集安人, 副教授, 从事数据仓库与数据挖掘等研究; 于戈(1962—), 男, 辽宁大连人, 教授, 博士生导师, 从事数据库理论和技术等研究.

2　数据仓库的结构及其描述

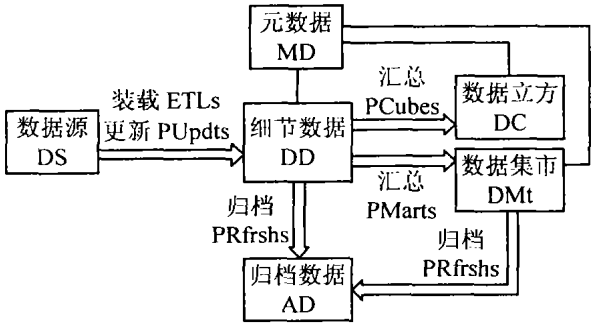

图 1　数据仓库的组成

数据仓库的组成如图 1 所示. 一个数据仓库 DW 可以表示为一个 6 元组

$$DW = \langle DS, DD, SD, AD, MD, PS \rangle.$$

DS 是数据仓库的数据来源, 它是由操作型系统中的数据库表 OST 和外部数据源 EDS (如电子表, 平文本等) 构成的集合, 即 DS= {OST, EDS, …}.

DD 是数据仓库中的细节数据集合. 细节数据是数据仓库的核心, 其数据直接来源于 DS, 或者是原始数据的存储, 或者是原始数据的聚合结果的存储. 它是整个企业范围模式一致的数据集合.

SD = {数据集市 DMt 的集合∪数据立方 DC 的集合}. 其中数据集市 DMt 用于企业中某一部门的分析, 它是由 DD 中的细节数据集成汇总而来, 而数据立方 DC 汇总的粒度相对较高, 用于完成某些探索性的分析, 如 OLAP 等. 因此它也被称为探索数据仓库[4].

AD 指归档数据. 随着时间的推移, 数据仓库中有许多数据很少被访问. 出于性能的原因以及辅存容量的限制, 需要将这些数据存储到近线或离线存储设备上.

MD 指数据仓库的元数据, 它是数据仓库系统的中枢. 元数据定义了数据仓库的体系结构、数据仓库的数据模型以及业务规则等, 为数据仓库的设计、维护和使用提供信息.

PS 指用于生成和管理数据仓库中各级数据的程序集合, 即

$$PS = \{ETLs \cup PMarts \cup PCubes \cup PUpdts \cup PRfrshs\}.$$

其中: ETLs 是数据抽取、转换和装载程序的集合, 而每个程序 ETL 是一个从 DS 到 DD 的映射, 即 ETL: DS→DD, 它完成将 DS 中的数据抽取出来, 进行转换后装入数据仓库的细节级数据库中; PMarts 是从细节数据生成数据集市的程序集合, 其中的每个程序 PMart 是一个从 DD 到 DMt 的映射, 即 PMart: DD→DMt; PCubes 是从细节数据生成数据立方的程序集合, 其中的每个程序 PCube 是一个从 DD 到 DC 的映射, 即 PCube: DD→DC; 数据仓库的更新是将数据源 DS 的更新传播到数据仓库的 DD 和 SD 部分; PRfrshs 将数据仓库中过时数据迁移到近线存储设备上, 实现从 DD, DC 和 DMt 到 AD 的映射, 即 PRfrsh: DD∪DC∪DMt→AD.

3　数据仓库设计方法

数据仓库系统从数据组织到支持的分析处理都与面向联机事务处理 (OLTP) 应用的数据库系统有较大差别, 这就决定了数据仓库系统的设计方法与传统的数据库系统的设计开发方法不同. 事实上, 企业级数据仓库的建设涉及面广, 是一项耗资巨大的系统工程, 因此需要完善的设计开发方法学作为指导.

本文总结并提出一个开发企业级数据仓库的框架, 其中包括 7 个步骤: 评估与规划、项目准备、需求分析与描述、数据仓库设计与实现、测试与完善、部署与培训和总结回顾. 其中第 3～7 个步骤是循环反复的过程, 即任何一个步骤发现问题, 都可以返回到上一步进行完善补充. 另外这个反复也指一个主题开发完成后, 进行下一个主题循环. 下面详细解释各个步骤的工作.

3.1　评估与规划

评估是进行任何项目必须经历的首要过程, 尤其是数据仓库项目. 对于耗资较大的数据仓库项目, 在作出投资决策之前必须进行充分的评估论证, 考察企业或机构是否有必要且有能力实施数据仓库项目; 然后要对数据仓库项目进行规划. 这个阶段的主要工作包括 4 个方面:

1) 进行数据仓库的可行性和必要性评估. 即结合单位的现状明确数据仓库建设的目标和任务. 另外, 要清楚数据仓库所面对的数据源所在系统和其中的数据状态, 并对相关的信息技术 (如数据源的数据库类型、工作平台、数据量、数据质量等) 进行评估. 通过对上述项目的评估, 核查建立数据仓库是否可行, 所建立的数据仓库是否是用户所希望的, 是否有不可逾越的障碍等. 另外, 要建立评定数据仓库项目是否成功的一些指标和基本原则.

2) 选择数据仓库的拓扑结构. 数据仓库的拓扑结构有 4 种[5]: 集中式企业级数据仓库, 独立型部门级数据集市, 分布式数据仓库, 数据仓库与数据集市

混合型. 大型企业一般选择数据仓库与数据集市的混合结构, 即从数据仓库导出数据集市, 以便为整个企业提供全局一致的数据视图.

3) 选择开发策略. 常用的开发策略有 3 种[4]: 自顶向下方法、自底向上方法、自顶向下和自底向上的联合方法. 数据集市的快速开发特性是解决这些企业需求的既快又节省投资的方案, 但是数据集市的简单堆积或连接在一起并不能构成企业级数据仓库[6]. 因此, 可利用自顶向下方法规划整个企业的数据仓库, 再利用自底向上方法快速开发数据集市.

4) 选择实现范围. 在总体规划确定了总方向和目标之后, 必须选定一个能够快速给企业带来效益的有限的实现范围, 即确定最初的实现范围. 对于企业而言, 可选择质量分析主题作为首选实现目标.

3.2 项目准备

该阶段主要工作包括收集分析企业OLTP 系统的结构和模型, 建立项目管理委员会和项目开发组.

分析OLTP 系统的目的是收集与现有OLTP 系统相关的文档资料, 找出源系统的整个或某部分概念模式或逻辑模式, 即收集元数据.

数据仓库项目是一个大工程, 需要很好的组织管理以及开发方和用户方的密切合作. 因此分清每个参与者的角色很关键. 关于这方面的详细论述见文献[7, 8].

3.3 需求分析与描述

本阶段需要设计者和数据仓库的最终用户合作收集并过滤用户的需求, 选择出用户分析处理所关心的事实, 并给出事实的描述、查询需求、报表需求和数据分析需求描述, 了解最终用户想进行的数据分析的类型, 如OLAP 或数据挖掘等. 尽管基于数据仓库的OLAP 或数据挖掘的需求很灵活, 没有固定模式, 但从上述各种不完全、甚至不明了的需求描述可以了解用户所关心并感兴趣的主要问题, 以及这些问题的解决需要什么样的信息等. 因此通过需求分析可以确定系统的边界, 找出数据仓库中的主要主题域.

通过本阶段的分析可以确定决策者所关心的事实、关于这些事实的度量指标、度量指标的粒度以及从哪些角度对这些指标进行分析. 进而, 可以确定探索数据仓库(数据集市或数据立方)的维及维层次.

3.4 数据仓库设计与实现

数据仓库的设计与实现包括概念设计、逻辑设计、物理设计和各种处理过程及应用接口设计与实现.

细节数据基本与数据源相对应, 所以细节数据中的数据集一般按第 3 范式设计. 而对于汇总数据, 尤其是数据立方, 则需根据分析主题的要求, 按多维方式进行组织. 下面讨论的方法主要是针对数据集市或数据立方.

OLTP 系统的 E/R 模型不适合为数据仓库的概念设计建立模型[9]. 因为 E/R 模型强调实体及它们之间的联系. Golfarelli 等[9]提出了维事实模型(DFM), 该模型可由 E/R 模型变换而成. 文献[10]将 E/R 模型和星型模式结合起来, 提出一种 Star-ER 模型. 文献[11, 12]则给出了从操作型系统的数据模型导出数据仓库或数据集市数据模型的方法.

这一阶段还要进行适当的粒度层次划分、合理的数据分割策略、关系模式的定义等. 另外可将一些查询视图实例化, 以便减少常用查询的响应时间.

物理设计所做的工作是确定数据的存储结构、索引策略、数据的存放位置和存储分配等[1]. 同时, 还要确定数据仓库的数据更新和净化策略. 数据仓库中数据的更新问题包括一致性要求、更新时间(即时的、周期的)、更新模式(在线、离线)和更新技术(重新计算、增量式)等. 数据仓库运行一段时间后就会产生"老化"数据, 清除老化数据的过程称为数据净化. 数据净化技术主要包括全部清除、有选择清除以及数据归档等.

数据仓库的各级模型设计完成后, 需要进行设计和编码, 用于生成和管理数据仓库中各级数据的程序集合 PS, 即设计和编写 ETLs, PMarts, PCubes, PUpdts 和 PRfrshs 程序代码, 并使用测试数据集对各个过程进行测试.

另外, 需要建立安全控制机制. 数据仓库系统中收集了企业和组织机构的重要敏感数据, 因此数据仓库的安全控制非常重要.

3.5 测试与完善

测试与完善阶段的目的是通过一个独立的测试组来确保数据仓库满足设计说明文档中的功能要求, 从而确保数据仓库的质量. 因此, 首先要建立测试数据仓库环境和元数据环境; 然后运行各个处理过程, 从数据质量、执行速度和安全性等方面评价它们的性能. 对于发现的问题和错误进行更改, 并记录有关的更改内容, 建立更改管理控制信息. 最后建立和提交集成测试文档.

3.6 部署与培训

测试完善之后, 需要生成数据仓库的正式版本

并发布运行, 即建立数据仓库的实际运行环境, 填入数据. 另外, 还要将数据仓库的组织以及数据的存取方法等传授给最终用户, 即培训用户. 培训的主要内容包括: 1) 向用户介绍数据仓库的全部情况(其中数据是重点, 不仅要介绍详尽的数据内容, 而且要介绍数据仓库系统是如何保障数据的质量、完整性和可靠性); 2) 告诉用户元数据的存储位置以及如何使用; 3) 数据仓库系统的前端数据存取工具的使用培训; 4) 数据仓库中数据的更新策略介绍; 5) 数据仓库安全规范的培训.

如果数据仓库管理员(DWA)不是数据仓库系统的开发者, 也需要培训DWA. 培训内容包括: 数据仓库的逻辑和物理模型、从OLTP 系统到数据仓库的数据流、全部的数据转换操作、所有元数据的存储位置和内容、数据装载和更新的策略、所有安全性问题及其测度以及所有程序文档资料的管理等.

3.7　总结回顾

总结回顾包含两个层面的工作: 一个是各个阶段的总结回顾, 称为进行中总结; 另一个是整个项目阶段性结束以及数据仓库运行一段时间后的总结回顾, 称为完成后总结.

项目进行中的总结主要是不断地总结回顾: 哪些地方可以做得更好, 业务部门对开发的支持是否到位, 双方如何合作得更好, 什么是业务部门见效最快的, 以及什么是开发部门见效最快的等, 以便在后续项目或下一个主题的开发中扬长避短. 另外, 当开发有了一定进展之后, 就要检查主题的范围选择是否恰当, 应参与的部门是否都积极主动地参与了工作, 有什么阶段性成果, 这些成果发布之后的用户反映如何.

项目完成后总结的主要内容包括: 数据仓库的建设是否对公司有所推进, 是否提高了公司的竞争优势, 投资回报率(ROI)是否达到了预计水平, 是否公司的其他部门可利用数据仓库获得效益, 是否得到未预料到的效益等.

上面给出了数据仓库开发的几个主要阶段. 这个开发方法充分利用了软件工程的思想, 既强调了系统开发的生命周期, 又兼顾了螺旋式开发方法(即先总体评估规划, 然后分步实施).

选择合适的数据仓库开发方法学对成功建设数据仓库尤为重要. 然而, 还有许多方面也会影响数据仓库项目的成功. 下面给出影响数据仓库系统成功开发的几个关键因素.

4　企业级数据仓库实施关键因素

数据仓库的建设不仅会给企业的各级决策者提供模式一致的数据, 而且会改变企业的文化, 即由以数据为中心转向以业务为中心, 由定性决策转向定量决策. 因此, 数据仓库的建设对企业的经营理念产生了巨大的冲击. 并不是每个企业都能成功地开发出有效的数据仓库, 事实上, 开发失败的案例要比成功的多[5]. 文献[5, 13]总结了企业成功地开发数据仓库的关键因素. 本文将这些因素归纳为 5 个方面, 即: 人的因素、需求因素、环境因素、技术支撑因素和质量因素.

4.1　人的因素

指企业管理者的支持和数据仓库潜在用户的广泛参与. 企业管理者的支持包括提供足够的可用资源, 同时将数据仓库作为企业度量指标和决策数据的唯一来源. 各级别的用户, 甚至包括高层领导, 必须提出他们的需求, 积极参与数据仓库的设计、开发和管理.

另外, 参与仓库设计、开发、实现和管理的人员必须理解决策信息的重要性, 能够分析和撰写业务需求文档, 全身心地投入数据仓库工程, 掌握和支配足够的资源, 具有项目开发和管理经验以及相应的知识, 熟悉开发工具和开发方法学等.

4.2　需求因素

指业务需求. 没有决定性的战略需求, 开发数据仓库注定要失败. 企业需求的最佳源泉是企业的战略规划和运营指标. 业务需求是企业信息体系结构和数据仓库体系结构设计的基础. 战略规划不仅为有效的管理提供指导, 也为企业内部改革提供了指导. 由于运营指标包括了报表(报告)的内容, 以及计算这些指标的数据来源, 所有运营指标报告综合起来便构成了数据仓库和企业战略信息系统的基础.

4.3　环境因素

指企业信息体系. 企业的信息体系是由企业的战略规划和企业数据体系(企业所需数据的完全范式化的数据模型)、信息系统体系(企业正在使用的用于生成、读取、更新企业数据的所有信息系统)以及企业技术体系(企业信息系统的硬件平台、操作系统、通信设施环境等)构成. 它是企业数据仓库系统建立和运行的环境. 因此, 在建设数据仓库之前必须搞清楚它将实现和运行的环境.

4.4　技术支撑因素

指数据仓库技术和体系结构设计. 数据仓库技术包括用户接口、数据仓库引擎、硬件平台、系统软

件和安全性问题. 数据仓库用户通过用户接口获得有用的信息. 选择用户接口的标准是所选择的数据仓库解决方案能够支持企业需求的变化、技术的提高和演进等. 选择的数据仓库引擎应能加载信息到仓库、实现存取控制(安全性)和支持多种接口工具集.

数据仓库体系结构是成功开发可伸缩数据仓库的关键因素. 企业数据仓库的体系结构设计应能反映企业的运营指标和业务需求. 数据仓库的数据模型、结构、组件和元数据应该基于企业内部信息需求, 而不是基于特殊的技术. 另外, 开发过程必须有成熟的开发方法学和相应的辅助工具加以支持.

4.5 质量因素

指信息质量. 数据仓库中的数据质量好坏是数据仓库工程成功的重要因素. 其数据必须正确、完整、及时、简洁和可理解[14]. 如果数据仓库中数据质量很差, 从中得到的信息(报表、模式)也就失去了辅助决策的意义. 如果用户对得到的结果产生了怀疑, 或发现数据仓库中包含质量较差的数据, 则用户就会很少或根本不再使用数据仓库中的数据. 这便意味数据仓库工程失败了. 如果仓库中包含错误的数据, 但用户还没有发现, 那么基于错误数据的决策将导致企业的经营失败. 这个因素与数据仓库体系结构中的 ETLs 密切相关. 因此, 在将数据导入数据仓库的过程中必须严格进行数据质量的检查和校对, 以保证进入数据仓库中数据的质量.

5 结 语

本文首先讨论了数据仓库的体系结构及其组成, 给出了数据仓库组成的形式化的描述, 这为数据仓库的构建提供了宏观的指导作用, 并明确了建立数据仓库要进行的主要工作. 然后, 根据该形式化的描述以及软件工程的思想, 提出了建立数据仓库的方法框架, 并对各个阶段的具体任务进行了论述. 最后, 总结了成功建立数据仓库的几个关键因素.

参考文献(References):

[1] Wu M C, Buchmann A P. Research issues in data warehousing [A]. *Proc of the German Database Conf* [C]. Ulm, 1997. 61-82.

[2] 王珊. 数据仓库技术与连机分析处理[M]. 北京: 科学出版社, 1998.

[3] Golfarelli M, Rizzi S. A methodological framework for data warehousing design[A]. *ACM 1st IntWorkshop on Data Warehousing and OLAP* [C]. Maryland, 1998. 3-9.

[4] Gill H S. 数据仓库—客户/服务器计算指南[M]. 王仲谋, 等译. 北京: 清华大学出版社, 1997.

[5] Gardner S R. Building the data warehouse[J]. *Communication of ACM*, 1998, 41(9): 52-60.

[6] Bontempo C, Zagelow G. The IBM data warehouse architecture[J]. *Communication of ACM*, 1998, 41(9): 38-48.

[7] Adelman S, Moss L T. 数据仓库项目管理[M]. 薛宇, 王剑锋译. 北京: 清华大学出版社, 2003. 187-205.

[8] SAS Institute Inc. Rapid warehousing methodology [TR]. NC: SAS Institute Inc, 2000. 18-26.

[9] Golfarelli M, Maio D, Rizzi S. Conceptual design of data warehouses from E/R schemas [A]. *Proc 31st Hawaii Int Conf on System Sciences*[C]. Hawain, 1998. (Ⅶ): 334-343.

[10] Tryfona N, Busborg F, Christiansen J G B. StarER: A conceptual model for data warehouse design [A]. *Int Workshop on Data Warehousing and OLAP* [C]. Kansas, 1999. 3-8.

[11] Moody D L, Kortink M A R. From enterprise models to dimensional models: A methodology for data warehouse and data mart design [A]. *Proc of the 2nd Int Workshop on Design and Management of Data Warehouses*[C]. Stockholm, 2000. 5: 1-12.

[12] Boehnlein M, Ende A U. Deriving initial data warehouse structures from the conceptual data models of the underlying operational information systems [A]. *Int Workshop on Data Warehousing and OLAP* [C]. Kansas, 1999. 15-21.

[13] Perkins A. Critical success factors for data warehouse engineering [EB/OL]. *DM Review*, 2002. http://www.dmreview.com.

[14] Redman T. The impact of poor data quality on the typical enterprise [J]. *Communication of ACM*, 1998, 41(2): 79-82.