

数据仓库的质量管理问题和方法^{*}

毛国君

(北京工业大学计算机学院 北京100022) (北京地区多媒体与智能软件重点实验室 北京100022)

The Problems and Methods of Quality Management in Data Warehouses

MAO Guo-Jun

(Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology, Beijing 100022)

(School of Computer Science, Beijing Polytechnic University, Beijing, 100022)

Email: maximmao@sohu.com

Abstract A data warehouse is often a large-scale information system for an enterprise, so its quality management is important and difficult. Recently, some researchers have studied the problems of quality management in data warehouses from different views, and achieved some good results. This paper will broadly introduce the concept, methods and techniques in quality management in data warehouses, and discuss the important quality factors in data warehouses in detail.

Keywords Data warehouse, Quality management

1. 引言

我们知道,在过去30年中,物理产品的制造过程已经被全面质量控制(TQC)或称全面质量管理(TQM)体系所改变。最早提出全面质量管理概念的是美国通用电气公司质量经理菲根堡姆,他在一本名为《全面质量管理》的著作中,提出“全面质量管理是为了能够在最经济的水平上并考虑到充分满足用户要求的条件下进行市场研究、设计、生产和服务,把企业各部门的研制质量、维持质量和提高质量活动构成为一体的有效体系”。20世纪60年代末,菲根堡姆的全面质量管理概念逐步被世界所接受,在实践上、理论上都得到了探索、总结和提高。特别是,在日本由于TQC的运用和普及,使产品的生产能力和信誉带来了前所未有的成功。随着国际贸易的迅速扩大,产品质量管理问题已经日趋国际化。为此,国际标准化组织(ISO)于1979年单独建立质量管理和质量保证技术委员会(TC176),负责制订质量管理的国际标准。1987年,我们熟悉的ISO9000~9004质量管理和质量保证系列标准正式发布。

进入20世纪90年代,软件产品的质量成为新的研究热点。的确,由于软件产品的特殊性,在其过程控制、量化管理以及指标标准化等质量管理相关方面难度更大。但是,随着软件产品向着大规模、可重用、国际化等方向的发展,实施全面质量管理也是必然的。事实上,软件质量管理的理论和方法也已经得到很好的研究。由美国卡耐基梅隆大学软件工程研究所(SEI)发布的CMM(软件能力成熟度模型),目前已在世界范围被广泛地推广和应用。

本文我们讨论数据仓库的质量管理问题。在这里,我们把数据仓库看作是一个信息产品。这不仅是因为数据仓库是软硬件的集成,更是因为它需要一个长期建设而又极易出现问题的信息系统。W. H. Inmon在80年代提出了数据仓库的概念,之后一些大的通讯公司和银行等开始尝试建立面向行业或全球组织的数据仓库系统。起初这类系统的成功率并不高,失败的百分比要比传统的信息系统高得多。但是,这并不是对数据仓库概念的否认,而是恰恰说明数据仓库建设中忽略了

质量控制问题。近几年,关于数据仓库建设面临的问题及对策和它的质量管理技术已经在许多文献中被讨论^[1~10]。DWQ是至今为止比较系统地对数据仓库质量问题进行研究的课题^[2,3]。该项目是由法国等4个国家研究机构联合立项的由欧洲ESPRIT资助的长远目标研究项目。它们指出了从语义层次上进行企业模型及质量管理的概念和方法,并在元数据层次中嵌入质量管理模型方面进行了有益的尝试^[2,3]。关于数据仓库中的质量管理问题的研究,目前有几个主要的切入点:(1)从数据仓库的设计入手,建立适合全方位质量控制的体系结构,如M. Jarke提出的基于概念(企业)、逻辑和物理三层模式的数据仓库构架^[3];(2)从软件生命周期观点,全过程控制数据仓库质量,如J. A. Rodero提出的按软件生命周期分阶段进行数据仓库审计(Audit)方法^[4];(3)按照数据仓库系统的工作机制和部件构成来制订质量控制标准,如J. A. Rodero对数据仓库的数据抽取、装载、存储等关键步骤和功能部件的质量标准和控制问题的研究^[5]。此外,大家公认的技术是:(1)量化控制信息,使数据仓库中的质量管理形式化,如M. Jarke应用的GQM(Goal-Question-Metric)方法来探讨数据仓库的质量控制标准^[3];(2)把质量控制信息和模型嵌入到数据仓库的元数据中,实现自动控制和审计功能。

在这篇文章中,我们将从数据仓库系统的层次模式和基本组成要素两个视点入手,讨论数据仓库的质量管理及其相应的控制目标。

2. 数据仓库系统的层次模式和质量控制

数据仓库质量是总体数据质量的问题。从本质上说,数据仓库的质量问题是一个主观的标准问题。不同层次的人员对数据仓库的质量有不同的要求,必须根据追求的目标来制定质量的评测、预测标准并加以实现。只有这样才能使系统化的质量管理成为可能。对于一个决策用户来说,他关心数据分析的能力以及由此而作出的决策的正确性、可靠性。对于一个数据仓库的设计者而言,他可能关心数据的组织形式及其合理性等。而一个数据仓库的管理员更关心数据的存储及高效存取问题。但是,质量控制标准又必须是客观的,即它们能准确

^{*} 本文研究得到国家自然科学基金(60173014)和北京市教委基金资助。毛国君 博士生,副教授,主要研究领域为数据挖掘和数据仓库等。

地反映企业对数据仓库的总体要求。

要正确理解数据仓库的质量管理问题,需要对数据仓库构造层次进一步细化和明确界限。数据仓库是一个面向企业级应用的数据管理体系和技术,从企业概念、逻辑和物理三个层次来理解它的构成及其相关的质量管理问题是科学的。在数据仓库的早期阶段,人们可能更关注逻辑和物理组织问题。但是,随着数据仓库规模的扩大,对企业模式(概念或语义层次)的刻画变得越来越重要,对应的质量控制问题必须提到议事日程上来。

流行的数据仓库构架大多都是把数据仓库看作是从源数据库抽取和集成得到的实视图。我们根据一些典型的数据仓库构架把这类数据仓库体系结构整理成图1。例如,著名的TSIMMIS^[6]和WHIPS^[7]等数据仓库项目都重点讨论了包装器(Wrapper)、监视器(Monitor)、集成器(Integrator)、仲裁器(Mediator)等功能部件的设计与实现问题,其贡献主要是在数据仓库的数据形成过程和集成方法等基础策略和技术上。

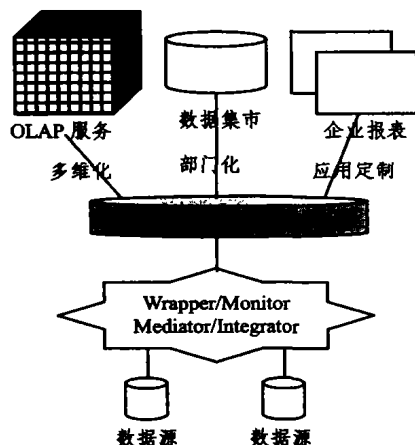


图1 典型的数据仓库结构示意图

从图1可以看出,这样的结构还是倾向于数据仓库本身的数据的逻辑和物理组织技术研究。但是随着基于数据仓库的企业应用的不断深入,暴露出潜在的缺欠。其中重要的恐怕就算是它对企业应用模式的缺乏。建立显式的企业概念模型层

次,能更好地支持企业应用和系统质量的控制^[2,3]。例如,AT&T的Manifold项目首先提出应该把领域的概念模型作为企业数据集成基础^[8]。另外,像前面提到的DWQ项目,提出了应该在概念、逻辑和物理三个层次上讨论数据仓库的设计和质量管理问题^[3]。综合这类研究,图2给出了一个基于概念、逻辑和物理三个层次的数据仓库架构。

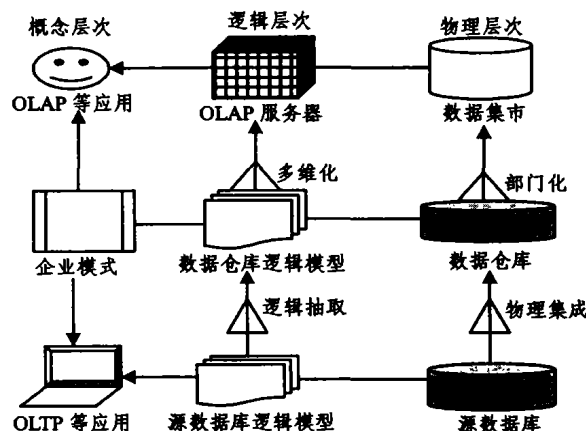


图2 三层模式的数据仓库结构示意图

依据图2这样的结构,我们可以更清楚地讨论数据仓库的质量管理问题和控制重点。简单地说,在概念或语义层次上,最终用户更关心企业模式的有效性、合理性和可解释性等。由于用户的应用是基于统一的企业处理模式进行的,所以对企业模式的评测和审计成为质量管理的重要内容。这就要求系统必须具有优秀的处理概念模式和数据的逻辑模型的转化能力。在数据仓库的逻辑层次上,定义了数据仓库的逻辑数据模型,对于特殊的应用可能需要进一步聚类数据成多维数据模型等。因此逻辑模型涉及到对企业应用的支持程度以及对源数据的集成能力。它应该更关注于逻辑模型本身的完整性、安全性、可用性以及对概念模型和物理模型的衔接能力。经过十几年的发展,数据仓库在其物理组织方面,特别是在大数据集的管理、性能优化、索引、快照等积累了许多经验。所以在物理

表1 数据仓库设计中的质量问题

指标	概念层次	逻辑层次	物理层次
正确性	概念模型与现实世界的吻合情况	描述现实世界实体的正确性	数据类型选择等的正确性
完整性	概念模型对企业处理与分析的覆盖能力; 概念模型表达商业知识的完整性	实体及其实体间的完整性; 操作的完整性	对源数据的抽取和集成程度
可存取性	概念模型到逻辑数据模型映射的正确性; 通过概念模型存取数据仓库数据的能力	逻辑数据模型映射到物理模型的正确性; 通过基于工具存取数据仓库数据的能力	形成存取数据仓库数据路径的优化能力
安全性	概念模型层次上可实施的安全措施与级别	逻辑模型层次上可实施的安全措施与级别	物理模型层次上可实施的安全措施与级别
可用性	概念模型的稳定性; 维护的时间和空间代价	逻辑模型的稳定性; 维护的时间和空间代价	索引、快照等优化措施的使用能力; 维护的时间和空间代价
可跟踪性	适应需求描述及其改变的能力	对概念模式改变的适应性	对逻辑模式改变的适应性
可解释性	概念模型的理解难度; 用户的使用难度; 对应文档的质量	逻辑模型的直观性;	管理的方便性
元数据	元数据对概念模型的表达能力; 对概念级质量管理模型的支持能力	元数据对逻辑模型及相关映射的表达能力; 对逻辑级质量管理模型的支持能力	元数据对物理模型及相关映射的表达能力; 对物理级质量管理模型的支持能力

层次上对数据的存取性能等因素考虑得更多。

下面我们就从数据仓库设计和基于数据仓库的信息系统的应用研制两个角度来刻画数据仓库系统中的质量管理相关问题。

2.1 数据仓库的设计质量

数据仓库的设计涉及到概念、逻辑和物理所有层次。为了建立一个高质量的数据仓库系统,必须首先弄清所有层次上可能出现的问题,并建立质量标准和实时评测,使问题及早发现和得到改进。表1给出了相应的工作要点。

2.2 软件实现质量

建立数据仓库的目的是为了应用。基于数据仓库的应用软件的质量也是完整的数据仓库系统(DWS)的一个重要的部分。的确,软件实现的质量问题不仅是基于数据仓库的应用所特有的问题。虽然数据仓库应用有它的特殊方面,但是我们可以借助于 ISO9126 标准来刻画它。表2给出了相应的工作要点。

表2 数据仓库系统的应用软件质量问题

指标	ISO9126	DWS
功能性	适宜性,精确性,协作性,安全性等	适合指定任务的程度;DW 的利用水平;与其它系统或模块的互动能力;安全使用数据仓库数据
可靠性	成熟性,容错性,可恢复性	软件的成熟水平;对数据仓库数据的容错能力;数据备份与恢复手段
可用性	理解能力,学习能力,操作能力	用户的接受程度;系统的智能程度;可用的辅助工具
软件效率	时间行为,资源行为	反映时间;传输时间;CPU/内存/硬盘等资源的利用情况
可维护性	分析能力,可改变性,稳定性,可测试性	错误报告水平及其修正和维护的代价控制;新业务的支持能力;系统化的测试与维护方法
便携性	自适应性,安装,一致性,替代能力	新环境下的安装时间;DW 重装后的适应性;数据一致性的保证措施

3. 数据仓库系统的组成要素和质量管理的

数据仓库系统是一个高冒险和高收益的工程。就管理而言,它的主要责任之一是保护企业所有的信息资产。为了实现它的任务和取得预想的结果,管理者必须提升内部质量控制系统以保护这些信息。从数据仓库的用户角度看,必须有满足最小质量需求的全过程保护措施。为了这个目的,建立独立信赖过程是必须的^[5]。J. A. Rodero 等提出了一个基于 COBIT (Control Objectives for Information and related Technology) 的数据仓库控制系统,并首次于1998年被 ISACA (Information Systems Audit and Control Association) 发表^[4]。它的核心是数据仓库评估控制系统。他们从数据仓库系统的基本构成要素入手,确定每个环节的控制目标,通过评估当前系统状况达到质量管理的目的。

为了更清楚地刻画对应问题,我们这里提供基于冒险评估的 (assessment of risks) 方法论的数据仓库质量管理方法。M Curtis 认为,数据仓库中的冒险是由于不当使用数据而构成损害或丢失的机会^[1]。我们可以从威胁项目的冒险因素入手,建立对应控制目标,并使这些危险减少到最低。下面我们按数据仓库形成、数据使用、人员三大类来阐述数据仓库可能

存在问题的地方及相应的质量管理问题。

3.1 数据仓库形成

这里是指对源数据、数据集市和数据仓库等存储体的管理,同时涉及到从源数据抽取、集成数据仓库的数据获取过程。源数据是数据仓库的数据供给,同时有担负着为操作用户的 OLTP 等服务。这些数据可能基于不同的操作型系统环境。数据集市和数据仓库是从源数据通过 Wrapper/Integrator 等功能部件获得的。从源数据抽取数据并装载到数据仓库中是建立数据仓库的关键性工作。因此,数据仓库的形成过程及其对应的存储体必须加以控制。通过确定对应的质量目标,可以保证得到一个理想的数据仓库实体。表3给出了对应的质量目标。

表3 数据仓库形成阶段对应的质量目标

对象	控制要点	质量目标
源数据	完整性,安全性,兼容性等	源数据本身质量:完整性、噪音数据量等; 管理源数据软件:安全性、兼容性、成熟性等
数据仓库	容错性,可恢复性,元数据,安全性,可用性等	数据仓库数据质量:完整性、覆盖能力、数据类型、稳定等; 管理数据仓库软件:容错、安全、备份恢复、优化技术等; 元数据:可扩展性等
数据仓库形成过程	抽取与过滤 清洗 同步 聚类 装载	抽取与过滤能力、时间、是否增量、易于理解等; 噪音数据的清除、空值处理、数据格式转换能力等; 具有插入数据的时态参考值(历史数据); 存储以前汇总数据、支持聚类模型、算法或工具的能力; 装载工具、临时的装载区等资源使用、安全计划、时间等

3.2 数据使用

这里是指对数据仓库数据利用能力的评估,这包括数据仓库的数据存取、分析、开采等方面。数据仓库的建立是为了使用。不管用户是否使用工具,数据仓库系统必须提供查询定义、信息搜索、汇总处理等基本功能,并为用户编程或使用分析工具提供接口。表4给出了对应的质量目标。

表4 数据仓库使用的质量目标

对象	控制要点	质量目标
数据存取	正确性、性能、安全性等	正确存取数据(如一致性、完整性); 存取数据的时间和资源占用; 持续存取数据和容错能力; 存取的安全性机制等
数据查询	定义、接口、标准、优化、显示等	界面的友好性; 接口的标准化; 内部优化机制
工具支持	对现有工具的支持、对新工具研制的支持等	支持 OLAP 工具; 支持数据挖掘工具; 支持灵活的报表或聚类分析或决策支持工具等

3.3 人员

这里的人员包括最终用户和系统分析、设计、开发、管理

等所有相关人员。它们的素质及其协作能力如何直接影响数据仓库系统的质量。实际上,从数据仓库概念提出开始,数据仓库系统的研究总是面对着三个主要层次的人员。第一是最终用户,他们是企业的业务分析者或决策者。他们是数据仓库系统框架下所有软件、工具、应用的最终受益者。他们利用数据仓库进行分析、决策、预测等工作。第二是辅助工具或应用软件的研制人员。应用软件或工具的设计和开发人员通过开发对应的软件或工具来满足最终用户的要求。他们可能以一定逻辑形式来理解数据仓库中的数据,并借助于合适的逻辑表示模式来重新组织数据仓库中的数据,以此来支持对数据仓库数据的高级分析工作。比如,一个 OLAP 工具,总是以多维结构来组织数据,为最终用户提供 OLAP 和数据挖掘等应用支撑。第三个层次是数据仓库管理人员 (DBA),它们必须更加关注数据的物理组织特性和性能。由于这些人员所扮演的角色不同,可以利用他们来评估系统和控制系统的进度等。表 5 给出了对应的质量目标。

表 5 数据仓库系统中人员相关的质量目标

对象	控制要点	质量目标
最终用户	使用、培训、安全、满意度等	利用数据仓库及其工具对分析、决策等工作的改善程度; 培训质量、难度、时间花费、教材; 权限等安全机制的使用; 整体满意度等
技术人员	项目主管、分析员、设计及开发人员等	各类人员能力、经验等评估(实施前); 队伍建设(如缺乏人员的补救措施); 技术人员的培训计划等
系统管理人员	DBA、硬件维护、网络管理等人员	数据仓库管理员水平评测; 硬件及网络维护人员水平评测; 应急事件处理能力等

结语 前面我们比较系统地介绍了数据仓库系统的质量管理及其控制目标。实际上,不论你从什么视点出发去进行数据仓库系统的质量控制,都必须要有周密而科学的计划保证。要很好地规划一个数据仓库系统的质量管理项目,下面的一些工作是必须的:

(1) 识别和文档化控制过程。这需要充分理解企业业务,包括需求、冒险和政策、组织结构、不同的角色和地位等。在此基础上确定质量控制策略、识别并制订质量控制指标、反复讨论并最终文档化。

(2) 实施控制和评估质量指标。在整个项目实施阶段,要收集和跟踪质量指标值,评测是否达到预定的质量控制目标。前面阶段的目标没有达到,不能进入下一阶段。对于潜在的冒险应尽早发现并加以控制。控制工作是一个连续的工作,整个过程要有一个合适的工作流程。控制工作应文档化,对任务检

查、结果、修订人、数据和时间等要明确加以注释。

(3) 及时完成审计报告。一旦得到现存控制和操作的所有信息,审计者必须得出当前控制的冒险估计。对每一个威胁因素,应该通过直接或间接证据数据加以量化,如发生概率和损害度等。这种评估是很困难的并且要以一个恰当的方式完成,它需要经验和分析能力。审计结果应该用报告形式给出。它包括审计的目标、覆盖时期和审计工作完成的特性和强度以及相关的结论和建议。报告应该送给相关受影响的管理者。

另外,本文没有涉及到质量管理的具体实现方式。正如前面所说的那样,建立合适的质量管理模型并把它嵌入到数据仓库的元数据中,是我们进一步需要探讨的问题。将来的工作也包括通过实际项目的应用,系统化提炼质量控制目标及其量化值,建立对应的原型系统等。

参考文献

- 1 Inmon W. Building the Data Warehouse. 2nd Edition, John Wiley & Sons Inc. 1993
- 2 Jarke M, Vassiliou Y. Foundations of data warehouse quality - A review of DWQ project. In: Proc. 2nd Intl. Conf. Information Quality, Cambridge, Mass. 1997
- 3 Jarke M, et al. Architecture and quality in data warehouses: An extended repository approach. Information System, 1999, 24(3): 229~253
- 4 Roderio J A, Piattini M G. Auditing Data Warehouses through the life-cycle. In: Proc. of the BDATOS'98 conf. Buenos Aires, Argentina. 1998. 9~20
- 5 Roderio J A, et al. The audit of the Data Warehouse Framework. In: Proc. of the Intl. Workshop on Design and Management of Data Warehouses (DMDW'99), Heidelberg, Germany, 1999. 14~18
- 6 Chawathe S, et al. The TSIMMIS project: Integration of heterogeneous information sources. In: Proc. IPSI Conf., Tokyo, Japan, 1994. 7~18
- 7 Hammer J, et al. The Stanford data warehouse project. Data Engineering, Special Issue Materialized Views on Data Warehousing, 1995, 18(2): 41~48
- 8 Levy A Y, et al. Data model and query evaluation in global information systems. Journal of Intelligent Information Systems, 1995, 5(2): 121~143
- 9 John W, et al. Maintenance of data cubes and summary tables in a warehouse. In: Proc. of the 1997 ACM SIGMOD Intl. Conf. on Management of Data, Peckman, J. M. (ed.), 1997. 100~111
- 10 Labio W, et al. Physical database design for data warehouses. In: Thirteen Intl. Conf. on Data Engineering, IEEE Computer Society, Birmingham, UK, 1997. 277~288
- 11 Curtis M, Joshi K. Internal Control Issues for Data Warehousing. Audit & Control Journal, 1997, 4