文章编号: 1001-0920(2001)02-0229-04

工业数据仓库设计方法及其在质量分析中的应用

嵇 晓[1], 鲍玉斌[2], 常 钊[1], 宋宝燕[2], 于 戈[2]
(1. 上海宝钢集团公司, 上海 201900; 2. 东北大学 信息科学与工程学院, 辽宁 沈阳 110004)

摘 要: 提出一种建立工业数据仓库的基本方法, 并结合某大型钢铁企业的具体情况, 给出一种数据仓库系统的实现方案, 讨论了数据仓库在企业产品质量分析中的应用。实践证明, 数据仓库可为企业的经营管理提供全面、准确的数据, 可在改进产品性能、提高产品质量方面发挥重要作用。

关键词: 数据仓库; 生产决策; 质量分析
中图分类号: TP 31 文献标识码: A

Design of Industrial Data Warehousing and Its Application in Quality Analysis

JI Xiao[1], *BAO Yu-bin*[2], *CHANG Zhao*[1], *SONG Bao-yan*[2], *YU Ge*[2]
(1. Shanghai Bao Steel Co Ltd, Shanghai 201900, China; 2. School of Information Science and Engineering, Northeastern University, Shenyang 110004, China)

Abstract: The methodology for building industrial data warehouse is presented to meet the production decision requirements in a large-scale iron-steel enterprise. An OLAP application for product quality analysis is described. The practical applications show that data warehouses can provide the complete and correct data for the decision-maker, and play an important role in enterprises, especially in improving product quality and performances.

Key words: data warehouse; production decision; quality analysis

1 引 言

经过多年的计算机应用实践, 企业已积累了大量、丰富、翔实的原始生产数据和各种业务数据, 这些数据真实地反映了企业主体和各种业务环境的经营动态。但由于缺乏集中存储和管理手段, 对如何充分有效地利用这些数据, 却一直没有很好的解决方法。因此, 充分利用企业积累的数据资源, 为企业的经营者和决策者提供决策支持, 是一项十分迫切的工作。数据仓库正是为决策者提供联机分析处理(如决策支持、数据挖掘)所需信息的仓储。它是面向主题的、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。国外关于数据仓库的应用项目已有许多。例如, 美国和日本的钢铁公司利用数据仓库和数据挖掘技术实现的 ISAP 系统, 研究分析产品性能规律并进行质量控制, 取得了显著的效果。

本文通过对某大型钢铁企业数据仓库的开发实践, 提出一种工业数据仓库系统的设计方法, 并以质

收稿日期: 2000-03-15; 修回日期: 2000-07-17
基金项目: 国家 863/CIMS 主题项目(863-511-946-004); 霍英东教育基金项目
作者简介: 嵇晓(1963—), 男, 上海人, 高级工程师, 从事数据仓库和数据挖掘研究; 于戈(1962—), 男, 辽宁大连人, 教授, 博士生导师, 从事数据库理论研究。

量分析为例介绍了数据仓库的应用。

2 数据仓库系统的设计方法

数据仓库从数据组织到支持的分析处理都与面向 OLTP 应用的数据库系统有较大差别。这便决定了数据仓库系统的设计方法不同于传统的数据库系统的设计开发方法。文献[1, 2]提出了数据仓库的设计方法, 本文则在此基础上, 结合取得的实践经验, 提出工业数据仓库设计的一般步骤如下:

2.1 规 划

规划是开发数据仓库系统必须经历的重要阶段。开发整个企业的数据仓库是一个长期的过程, 需要统筹规划, 尤其是采用自底向上的开发策略, 更需要规划企业全局的数据模式。该阶段的主要工作包括:

1) 选择数据仓库的拓扑结构: 数据仓库的拓扑结构有 4 种, 即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[3]。大型企业一般选择数据仓库和数据集市的混合结构。因为大型企业数据丰富、物流复杂, 分析所用数据来源广泛, 利用集中式数据仓库可为企业提供清洁而模式一致的数据, 便于实现复杂的需要多部门数据的分析处理。

2) 选择开发策略: 常用的开发策略有 3 种, 即自顶向下方法、自底向上方法、自顶向下和自底向上的联合方法[3]。数据集市的快速开发特性是解决企业需求的既快又节省投资的方案, 但数据集市的简单堆积或将数据集市连接在一起并不能构成企业级数据仓库[4]。因为如果没有全局的组织规划和数据集市建模, 设计时不考虑跨部门的信息需求, 则所建立的数据集市之间将产生零碎且模式不一致的数据, 这将妨碍将来需要跨部门信息的应用开发。因此, 可利用自顶向下的方法规划整个企业的数据仓库, 利用自底向上的方法快速开发数据集市。

3) 选择实现范围: 在总体规划时已确定总方向和目标后, 必须选定一个能快速为企业带来效益的有限的实现范围, 即确定最初的实现范围。对工业企业而言, 可选择质量分析主题作为首选实现目标。

2.2 OLTP 系统分析

数据仓库中的数据主要来源于 OLTP 系统的数据库, 是对原有数据库进行集成和重组而形成的数据集合。因此, 弄清 OLTP 系统中实体间的关系, 对找出主题和事实是非常必要的。事实是决策制定过程中感兴趣的概念, 与企业中动态发生的事件相对应, 在 E/R 图中事实可以是实体或关系。该阶段的目的是收集与现有的 OLTP 系统相关的文档资料, 找出原系统的整个或某部分概念模式或逻辑模式, 即收集元数据。元数据是“关于数据的数据”, 它包括业务元数据和技术元数据。业务元数据是最终用户使用的一类元数据, 包括有关应用细节的文档、业务概念和术语、有关预定义查询和报告的详细情况、上下文信息等。技术元数据是开发与维护系统的数据库管理者和应用程序开发者使用的一类元数据, 主要包括数据源、数据仓库和数据转换规则等数据字典。

2.3 需求分析与描述

该阶段要求设计者和数据仓库的最终用户合作收集和过滤用户的需求, 选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。查询需求是用最终用户术语描述的业务查询样本。不同的最终用户提出的查询可能不同, 如销售部门会提出“前 6 个月哪个地区一直完成得最好”, 而技术部门会提出“半年来工艺参数的控制水平如何”等。

2.4 数据仓库设计与实现

数据仓库的设计包括以下内容:

1) 概念设计: OLTP 系统的 E/R 模型不宜作为数据仓库概念设计模型[1]。E/R 模型强调实体及它们之间的关系, 而 OLAP 系统则通常先确定分析主题, 然后确定它们的维及维层次。事实可从 E/R 模型的实体和关系中抽取, 因此应充分利用 E/R 模型提供的信息来设计数据仓库的概念模型。

2) 逻辑设计: 逻辑设计是以概念设计得到的维模型、需求描述等信息, 产生能尽量减小响应时间的数据仓库模式。数据仓库的逻辑模型可以是关系的, 也可以是多维的。多维模型主要以星型模式或雪片模式为代表, 并可影射到关系模式。另外, 逻辑设计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、查询视图实例化等。

3) 物理设计: 物理设计工作包括: 确定数据的存储结构、索引策略、存储位置和存储分配等。

2.5 数据仓库的生成与维护

数据仓库的维护包括: 数据抽取、清洗、装入、更新、净化、元数据管理及性能监视。数据抽取是从数据源读取数据。源数据读出后还需对其进行清洗, 以便除去不真实数据, 再经模式集成、语义转换、聚合运算后装入数据仓库。这一过程称为数据仓库的

初始生成。此后的源数据修改应反映在数据仓库中, 该过程称为数据仓库的更新与净化。数据仓库中数据的更新包括: 一致性要求、更新时间(即时的、周期的)、更新模式(在线、离线)、更新技术(重新计算、增量式)等。数据仓库只能保留一定时间段的数据, 它在运行一段时间后将产生"老化"数据, 清除老化数据的过程称为数据净化。数据净化技术主要有全部清除、有选择清除和数据归档等。

3　钢铁技术质量数据仓库的实现

在钢铁企业技术质量数据仓库建设中, 我们设计了一种混合型体系结构。系统中数据抽取程序从数据源读取数据, 再经数据转换处理后形成各主题下的原始数据集, 原始数据集与数据源系统中的数据集相对应。在原始数据的基础上生成汇总表、数据集市和信息集市。汇总表中定义了数据汇总处理的维、层次及事实等, 即数据立方。数据集市设在数据仓库内, 而不在主题中, 即数据集市可跨主题进行数据重组。信息集市是数据仓库中数据处理后产生的结果集合, 可以是一些报表、图形或分析结果。为提高分析速度, 预先将一些需经反复进行的相对固定的分析决策的结果存储在数据仓库中, 当决策者需要这些信息时, 能立即从信息集市中得到。在数据集市和汇总表上可进行查询、OLAP 和数据挖掘等分析工作。元数据仓库用于收集数据仓库建设中每一步的元数据。

下面介绍利用技术质量数据仓库进行 DI 材质分析的应用情况。技术质量数据仓库收集了从工厂生产一线采集的实绩数据, 包括炼钢、热轧、冷轧和电炉等工艺环节, 以及来自业务处理各阶段的数据, 如用户合同、质量设计后的生产合同、材料设计及质量设计等。以业务为中心建立起数百个主题, 例如热轧产品质量分析、冷轧产品质量分析、DI 材质分析等。

DI 材的技术质量指标主要有: 调质度、屈服强度、拉伸强度和各向异性等。材质各向异性(Δr 值)是指材质在 0°、45°和 90°等各方向的均匀性, 它的好坏将直接影响制罐的成罐率。据统计, 投产初期的一段时期内, 因 DI 材的 Δr 值超标而改判的比例平均高达 10%, 极大地制约了 DI 材成材率的提高, 影响了制罐使用的稳定性。另外, 由于 Δr 值不良和 YP 偏高导致冲罐缺陷而隔离的钢卷占所有隔离卷的 75%以上。针对这些问题, 我们利用技术质量数据仓库中关于 DI 材的主题数据, 研究分析生产过程中工艺控制参数和材质特性水平及其关系, 以便找出影响 DI 材产品性能的原因和解决方法。

改善材质各向异性的主要手段是提高成分、热轧温度、冷轧变形量、退火温度和速度等关键工艺参数的控制水平。其中成分是最关键的因素之一。成分碳(C)与塑性应变比(r)和 Δr 值有明显的相关性。随着 C 含量的增加, r 和 Δr 值减小; 当 C 含量在 0.03%以上时, Δr 值完全可控制在 0.2 以下而满足制罐要求。

下面利用金相分析和拉伸试验来验证相关性分析得出的结论。当 C 含量为 0.021%时, 从 0°、90°和 45°方向的断面金相组织比较分析可知: 90°方向的晶粒明显粗大, 0°其次, 45°最小。对应的 r 值分别为 1.47, 1.52 和 1.17, 也证实了各个方向上的晶粒延展性差异较大。由公式 $\Delta r = (r_{0°} + r_{90°} - 2r_{45°})/2$ 计算知, 此时 $\Delta r = 0.32$, 严重超标(≤ 0.26)。但通过分析可知, 抑制 0°和 90°方向晶粒过分粗大, 缩小 $r_{0°}$和 $r_{90°}$与 $r_{45°}$的差距, 可促使 $\Delta r \to 0$, 即材质均匀化。当 C 含量为 0.029%时, 由 0°、90°和 45°方向的断面金相组织比较分析可知, 各方向的晶粒大小没有明显差异。拉伸结果表明 $r_{0°} = 1.0$, $r_{90°} = 1.1$, $r_{45°} = 1.03$, 显然差异不再存在, 此时 $\Delta r = 0.028$, 材质趋向于理想的均匀化。由此可认为 C 含量增加, 能有效地抑制晶粒粗大, 减小晶粒的差距, 达到改善材质均匀化的目的。

在充分保证其它力学性能稳定控制的前提下, 对成分 C 含量进行了适当上调。调整后 C 含量平均值已达 0.03%(见图 1)。相应地, Δr 值控制情况也大为改善(见图 2), Δr 平均值仅为 0.05, 比较集中地分布在目标值0附近; 工序能力指数(C_{pk}值)达 1.3, 控制能力充分。成分 C 含量调整后, 材质各向异性已处于较为理想的稳定控制状态。

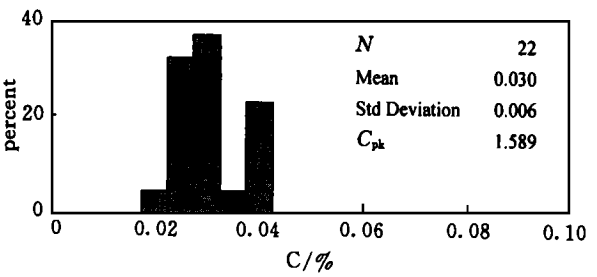

图 1　调整后 C 成分控制直方图

计算 C_{pk} 的下限值为 0.00, 上限值为 0.10

利用质量数据仓库中DI材质分析主题数据, 经

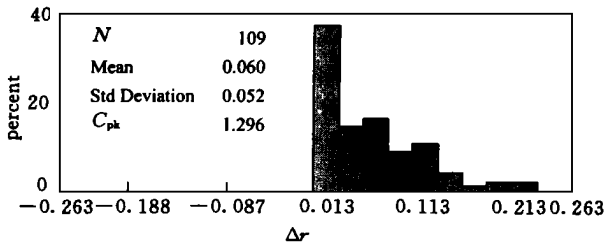

图 2 Δr 控制情况

计算 C_{pk} 的下限值为 − 0.260, 上限值为 0.260

过长时间的跟踪分析, 在逐月进行SPC分析的同时计算出 C_{pk}, 利用多元回归分析和方差分析, 找出 Δr 与 C 含量之间的关系。根据得到的关系及实验分析验证, 调整工艺参数控制, 使 Δr 值的控制情况大为改善。改善措施实施后的第一个月, Δr 值不符改判率就已控制在目标($\leq$ 5%) 以内, 且控制水平逐步提高。两个月后Δr值超标改判率为0, C_{pk} 值为 1.65, 达到历史最好水平。

4 结 语

在企业内大范围使用数据仓库技术, 能有效地管理庞大的信息资源, 更重要的是改变以往数据库技术 "以数据为中心 "的理念, 强调 "以主题、业务分析为中心, 以决策为目的 "的思想, 在质量分析工作中应用数据仓库, 可大大提高质量管理工作的效率, 管理大量洁净的模式一致的数据, 为工业企业充分利用数据资源提供了途径。基于数据仓库的应用不但能节省用户的大量时间, 而且可提供既准确又有权威性的统计分析结果, 为及时掌握产品的质量趋势提供有效的保证, 为企业的生产经营决策提供有力的支持。

参考文献:

[1] M C Wu, A P Buchmann. Research issues in data warehousing[A]. Proc of BTW'97[C]. Ulm, 1997. 61-68.

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

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

(上接第 228 页)

4 结 论

利用现场总线构成的 $N + M$ 容错系统, 可以方便地实现基本控制单元与备用单元之间的无扰热切换和各种故障容错策略。与传统的方法相比, 该系统具有结构简单、可靠性高、柔性好等特点, 对提高复杂系统的可靠性很有成效。本文阐述的基于CAN总线的机器人化遥控铲掘机 $N + M$ 热切换容错系统, 不仅适用于 RRS 系统, 而且适用于其它工业控制系统。

参考文献:

[1] 阳宪惠. 现场总线技术及应用[M]. 北京: 清华大学出版社, 1999.

[2] Losq J. Influence of fault detection and switching mechanisms on the reliability of stand-by system[A]. Proc 5th Int Symp Fault-tolerant Computing[C]. American: IEEE Computer Acadermy Fault-tolerant Technology Committee, 1975. 81-86.

[3] 邬宽明. CAN 总线原理和应用系统设计. 北京: 北京航空航天大学出版社, 1996.

[4] 路同浚, 孟宪超, 吴平川, 等. 机器人化炉渣铲掘机的研究[J]. 机器人, 1999, 20(4): 317-321.