学校编码：10384　　分类号_____密级__
学　　号：24320071151842　　UDC __

厦门大学

硕 士 学 位 论 文

数据仓库中数据质量管理的研究与应用

Research and Application of Data Quality Management in Data Warehouse

赵 彦 泓

指导教师姓名：王 备 战 教授

专 业 名 称：计算机软件与理论

论文提交日期：２０１０ 年 ５ 月

论文答辩时间：２０１０ 年 月

学位授予日期：２０１０ 年 月

答辩委员会主席：_____

评 阅 人：_____

2010 年 5 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　月　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

建设企业级数据仓库是目前企业信息化的发展趋势。建立数据仓库，通过分析企业数据和客户信息，基于数据为企业提供决策支持能够有效提高企业的市场竞争力。随着企业数据仓库的深入应用，数据质量问题渐渐暴露出来，而数据仓库项目的成功与否直接依赖于数据质量的好坏。高质量的数据是企业能够制定良好决策的基础，而低质量的数据将严重降低基于数据仓库做出的决策的可信度。如何控制数据仓库的数据质量是目前数据仓库建设中日益突出的问题。

论文的主要工作表现在以下方面：

1. 归纳了数据质量问题的研究背景、研究现状以及数据仓库和数据质量的相关概念，阐述了保证数据仓库数据质量的一般方法。
2. 研究了数据仓库中数据质量检查技术、原则和标准，基于项目经验总结归纳了数据仓库中常见的数据质量问题，提出了改善数据仓库数据质量的基本步骤和方法。
3. 探讨了元数据（Metadata）对于数据质量的重要性以及基于元数据控制数据仓库数据质量的方法。
4. 将研究成果应用到实际项目中，改善了数据仓库的数据质量。

关键字：数据仓库；数据质量；元数据

Abstract

To set up Data Warehouse of enterprises is in accordance with the need of developing their information technology. Through the establishment of Data Warehouse and the analysis of enterprise data and client data, we can provide data-based decision supports to enhance the competitive power of an enterprise effectively. As the enterprise Data Warehouse is applied more and more widely, the problem of data quality also emerges, which exerts a huge influence on the construction of a Data Warehouse project. High-quality data lays the foundation for good company decisions, while low-quality data will affect the reliability of decisions based on Data Warehouse. How to control the data quality has become a vital issue when constructing Data Warehouse.

This dissertation makes a contribution in the following aspects:

1. It introduces the research background and related researches of data quality, elaborates on its concept, explores the general methods adopted to guarantee the data quality.
2. It exerts efforts to explore techniques, principles and standards of examining data quality, make a summary of problems that frequently happen in Data Warehouse based on project experience, and bring forth with methods to improve data quality.
3. It discusses the important role that metadata plays to influence data quality and the means through which it controls data quality.
4. It applies research findings in actual projects and improves the data quality of Data Warehouse.

Key Words: Data Warehouse; Data Quality; Metadata

目 录

Contents

第一章　绪论

1.1 研究背景

过去的几十年里，世界经济的发展已经从产业经济时代进入了信息经济时代，数据在企业发展过程中扮演的角色越来越重要，逐渐成为企业重要的战略资源。知识工作者已经替代工厂工人成为这个时代劳动力的主力军，企业对信息的吸收反馈能力直接影响到企业的竞争力。为了给企业提供更好的数据服务和决策支持，数据仓库应运而生。数据仓库的产生使数据不再简单的用于检索，企业可以用它来分析整个企业的运行状况以及未来的发展趋势，它的出现改变了整个企业信息化格局。

建立企业数据仓库对于企业来说意义重大：

首先，它可以提高企业的市场竞争力，使企业能够提供更好的客户服务。在过去的计划经济条件下，国内的企业很少去关心市场的概念，当时的市场是国家划分好的，客户也没有太多的选择权，由于市场竞争压力小，国内的企业在提高市场竞争力方面的经验和适应能力也相对较薄弱；国外的企业一直处于激烈的市场竞争环境下，在提高市场竞争力以及对客户数据进行分析方面积累了大量的宝贵经验。建立企业数据仓库，为市场营销和客户分析提供最基本的数据源和辅助工具，是企业提高自身市场竞争力和客户服务水平的关键。

其次，它能提高企业的资产质量，防范金融风险。企业资产的保值、升值是企业发展最基本的保障，不良资产的产生很大程度上源于企业资金管理体制的不完善，不能有效利用企业的信息数据将导致银行的管理决策层对信息的反应滞后，无法有效预测及防范风险，造成企业资产的大量流失，进而影响企业的健康发展。有效利用数据仓库技术，建立企业数据仓库，对企业的数据信息进行有效管理，深入了解企业的客户信息，可以从企业的整体出发，综合管理企业资产，为企业资产的优化提供良好的解决方案，有效控制风险，提高企业的资产质量和利润率。

最后，它能够提高企业的经营管理水平，降低成本，提高企业的效率。企业要想在激烈的市场竞争中取得优势，必须为企业建立科学的管理决策机制，这就

需要科学化管理决策工具的支持。建立企业数据仓库，可以实现对企业信息的有效管理，分析产品、部门、机构的利润和成本，通过加强成本管理来增加企业的效益；同时，改进企业各级部门的管理、控制和协作的手段，可以使整个企业的经营管理更加科学、有效和规范。

但是目前许多企业在数据仓库项目的实施过程中，经常忽略一个关乎数据仓库项目能否成功的重大问题，即数据仓库的数据质量问题。数据仓库的数据来自许多不同的数据源，中间还经过数据 ETL（抽取、转换、装载）等过程，难免会出现一些数据问题，而这些问题会严重影响数据仓库的数据质量。高质量的决策依赖于高质量的数据，面对复杂的企业数据环境，如何有效控制数据仓库的数据质量是数据仓库建设过程中一个至关重要的问题，也是目前困扰企业经营管理决策的一个问题。

1.2 研究现状

国内对数据仓库数据质量问题的研究起步较晚，而国外在这方面已经做了许多开创性研究，并且已经取得了一定的成果。随着国内企业信息化进程的推进，数据质量问题也受到越来越多的关注，逐渐展开了许多以数据质量问题为中心的研究。

DWQ 是迄今为止比较系统地对数据仓库的数据质量问题进行研究的项目[1]，它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[2-5]。除此之外，目前国内外关于数据质量问题的研究成果主要有：

1. 麻省理工大学 Richard.Y.Wang 教授领导的 TDQM(Total Data Quality Management)研究小组提出的全面数据质量管理方法和数据产品质量评估方法[6]。
2. 复旦大学数据质量研究小组提出的一个可扩展数据清洗框架的定义[7]。
3. 北京大学数据质量研究小组提出的用六元组描述的数据质量评估模型[8]。

关于数据仓库中的数据质量问题的研究，目前有几个主要的切入点:

1. 从数据仓库系统的工作原理出发制定相应的数据质量控制标准，比如

J.A.Rodero 对数据仓库的数据抽取、装载、存储等关键步骤的数据质量标准和控制问题的研究[9]。

2. 从软件生命周期出发，对数据仓库建设的全过程实行质量控制，如 J.A.Rodero 提出的按照软件生命周期分阶段对数据仓库进行审计的方法[10]。
3. 从数据仓库本身的设计入手[11]，建立适合对数据质量进行全方位控制的数据仓库体系结构，如 M.Jarke 提出的基于概念、逻辑和物理三层模式的数据仓库体系结构[12]。

对于数据仓库中数据质量控制问题，目前业内有两种公认的技术：

1. 量化数据质量控制信息，使数据仓库中的数据质量管理形式化，比如 M.Jarke 应用的目标问题矩阵方法来探讨数据仓库的质量控制标准[13]。
2. 把数据质量控制信息和模型嵌入到数据仓库的元数据中，实现数据质量的自动控制和审计[14]。

目前对数据仓库中数据质量问题研究主要涉及到的技术手段有以下几种：

1. 隐藏规则的挖掘与异常发现：挖掘数据仓库中满足某些质量指标的所有规则，发现与这些规则不一致的异常数据，并进行适当的处理。
2. 重复或相似记录检测：重复或相似记录是数据质量问题分析的难点之一，通常是在排序基础上对数据进行相似性分析。
3. 孤立点检测：所谓孤立点是指在数据源中与众不同的数据，它们通常具有与常规数据模式显著不同的数据模式，这可能是度量或执行错误导致的，也可能是固有数据变异的结果。
4. 缺失值处理：目前对缺失值进行处理的方法主要有常量替代法、平均值替代法、最常见值替代法和估算值替代法等。

在数据仓库的发展过程中，关于数据质量的工具主要可分为以下几类[15-17]：

1. 协调不一致的数据并确定数据的完整性；
2. 识别与数据域不对应的数据；
3. 协调参照完整性的问题：大多数数据库管理系统都能对数据的参照完整性进行检查，即根据主键和外键的定义检查各表间的关系；
4. 评价现有数据的质量：这类工具主要对数据进行值域检查；

5. 数据重组：对数据进行重组，改进性能，提高数据的完整性，更好的支持生产应用程序，为决策支持做准备；
6. 提供度量数据质量的结构；
7. 数据字典/模式库：一个数据字典/模式库是质量数据的重要组成部分，通过它可以充分理解数据的涵义、数据的来源、数据的完整定义以及数据质量的象征意义。

国内在构建企业数据仓库上的起步较晚，在技术方面还处于学习和发展阶段，成功实施和应用数据仓库的案例还不是很多，导致这些巨资投入的数据仓库项目达不到预期目标而备受谴责和争议，而究其根本，很多数据仓库项目失败的重要原因是数据质量问题。在企业复杂的数据环境中，由于源系统众多，数据量庞大，各源系统间的数据定义和管理标准不一致等原因，尽管在数据转移到数据仓库环境的过程中会进行数据清洗转换等处理，但缺乏有效的元数据管理，缺少科学的数据控管，数据质量将难以得到保证，许多企业的数据仓库项目都是由于数据质量的问题造成了很严重的数据错误，导致客户对数据仓库的数据信任度下降。

作为决策支持系统的基础，数据仓库必须提供高质量的数据和服务。在数据仓库的设计和运行过程中，必须时刻注意保持数据的一致性、完整性、准确性、可用性、以及良好的系统性能等。如何检测并排除潜藏在数据仓库中的数据错误，以保证数仓库中数据的质量，进而为正确的决策打下坚实的基础，这是在建设数据仓库时必须重点考虑的核心问题，也是数据仓库项目成败的关键。

1.3 论文的主要工作

企业数据仓库的建设是一个长期复杂的工程，改善企业数据仓库的数据质量更是一个持续不断的过程，需要有一套科学有效的数据质量改善方法以及技术人员和业务人员的通力配合。

论文基于某银行数据仓库项目经验，对数据质量问题处理方法进行研究，提出了改善数据仓库数据质量的步骤和方法，并将研究成果应用到实际项目中。

论文的结构安排如下：

第一章 主要归纳了数据质量问题的研究背景、研究现状以及论文的主要研