元数据管理平台总体架构设计研究

Research on architecture of metadata management platform

杨鸿宾　宋 明　（北京邮电大学经济管理学院 北京 100876）

摘　要：元数据管理解决的问题就是如何把业务系统中的数据分门别类地进行管理，并建立数据与数据之间的关系，为数据仓库的数据质量监控提供基础素材。元数据管理指管理数据仓库系统的元数据，它贯穿数据仓库系统的各个环节，并实现系统的各个处理单元由元数据驱动。

关键词：元数据管理 数据质量管理 数据仓库 商业智能

1 引言

元数据常见的定义是："关于数据的数据"。更准确一点说：元数据是描述流程、信息和对象的数据。这些描述涉及像技术属性（例如，结构和行为）这样的特征、业务定义（包括字典和分类法）以及操作特征（如活动指标和使用历史）。数据仓库项目收集到的海量数据中隐藏着大量珍贵的信息，但也同时提高了系统的数据管理难度。一方面难以对这些数据进行有效解释，缺乏对业务流程执行的管理；另一方面各部门数据与数据整合的难度也不断加大，影响到了系统的数据质量。元数据管理成为解决诸多问题要抓住的一个"精髓"要素。

2 元数据的分类

元数据贯穿数据仓库系统数据"流动"的全过程，主要包括数据源元数据、数据采集元数据、数据仓库存储元数据、数据集市元数据、应用服务层元数据和门户管理元数据。根据元数据用途及针对使用角色的不同，分为技术元数据、业务元数据和管理元数据三类：

（1）技术元数据。面向运维技术人员，偏重数据结构和数据处理细节方面的技术化描述，是用于开发和维护经营分析的基本信息；

（2）业务元数据。面向业务分析人员，是对经营分析的数据和处理规则的业务化描述，主要包括业务规则、业务术语、指标业务口径、信息分类等；

（3）管理元数据。面向数据仓库运维管理人员，是对数据仓库运维管理相关信息的描述，主要包括管理流程、人员职责、工作内容分配描述等信息。

元数据贯穿数据仓库系统数据"流动"的始终，实施元数据的集中管理可以提供一个集中的元数据全局视图，对数据的组成、转换以及来龙去脉有效地进行数据质量的管理。

3 元数据管理平台的意义

从技术理论上讲，元数据涉及到商业智能中的数据仓库、ETL、联机分析处理、数据挖掘、前端展现等多方面内容，元数据贯穿项目的始终。从技术实现上讲，元数据分布在仓库的不同组件中，业务规则和技术元数据是分离的，而且由不同系统以不同格式保存且用户界面不同，不利于业务人员和技术人员对于元数据的管理和使用。

另一方面，多数用户完成了数据仓库建设中的数据整合工作，数据质量的保障和全面提升成为制约系统应用、推广和后续建设的关键，良好的数据质量是数据仓库价值体现的基础。而根据从数据仓库的系统中获得的数据做出智能决策和采取信息化行动时，分析人员和决策者需要知道自己的需要与经营分析系统中数据的关系。建立系统元数据的管理平台，使得技术人员和业务人员可以统一地对数据仓库系统中的元数据进行管理和监督以及探查。建设完善的数据质量管理体系和规范管理流程，元数据是数据质量管理的重要组成。

元数据管理指管理数据仓库系统的元数据，它贯穿数据仓库系统的各个环节，并实现系统的各个处理

单元由元数据驱动。参见图1。

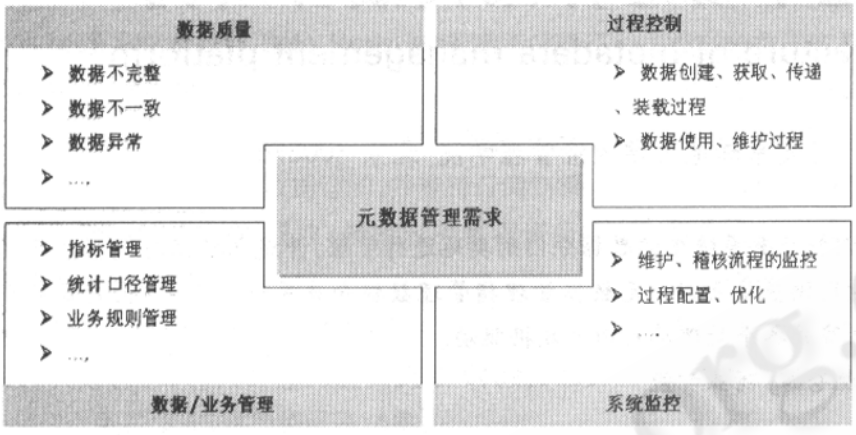

图 1　元数据管理的需求

4　元数据管理平台的总体架构

4.1　基本框架(参见图2)

元数据管理系统以元数据库为中心展开,各子系统元数据经过转换工具或者 API 转换为 XMI 文件或者调用元数据库的 CORBA IDL 接口将元数据导入元数据库。同时,元数据管理工具也可以通过 CORBA IDL 或者 XMI 文件的形式将元数据库中的元数据内容返回。

元数据库提供 CORBA IDL 接口为元数据管理工具提供访问和管理接口,此接口与各子系统和元数据库交互的 CORBA IDL 接口是相同的,这是由 CWM 标准本身决定的,这也使得元数据管理工具有能力直接访问某些支持 CWM 标准的数据仓库系统。

4.2　功能框架(参见图3)

元数据平台功能框架分为元数据源层、元数据获取层、元数据存储层、元数据管理层和元数据访问层。元数据源层包括元数据的各个源系统;元数据抽取层中的连接桥(或称适配器)实现元数据源层元数据的抽取;元数据抽取层抽取出的元数据存

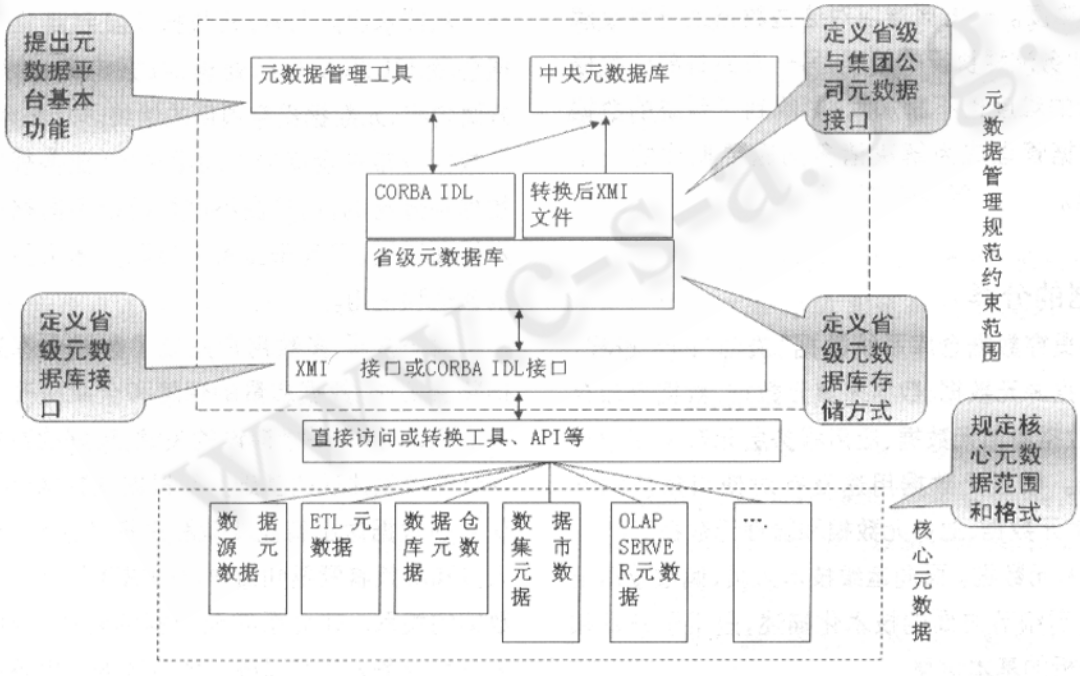

图 2　元数据管理和数据仓库的交互关系

元数据管理系统以元数据库为中心展开,各子系统元数据经过转换工具或者 API 转换为 XMI 文件或者调用元数据库的 CORBA IDL 接口将元数据导入元数据储在元数据存储层中的元数据库中;元数据管理层提供元数据访问、分析、导入、导出等功能,供元数据管理工具前端、二级数据仓库系统以及中央元数据抽取服

务器使用。

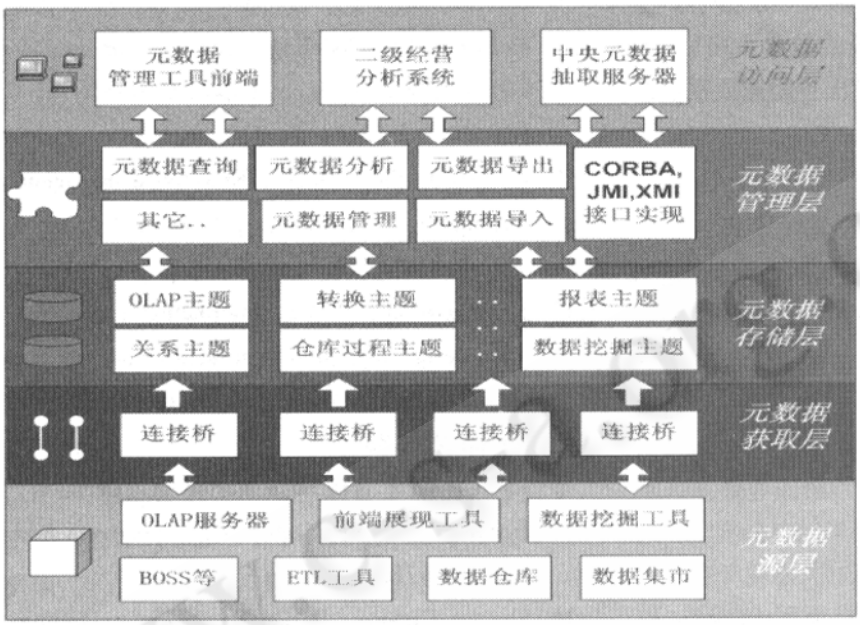

图 3　元数据管理平台层次关系

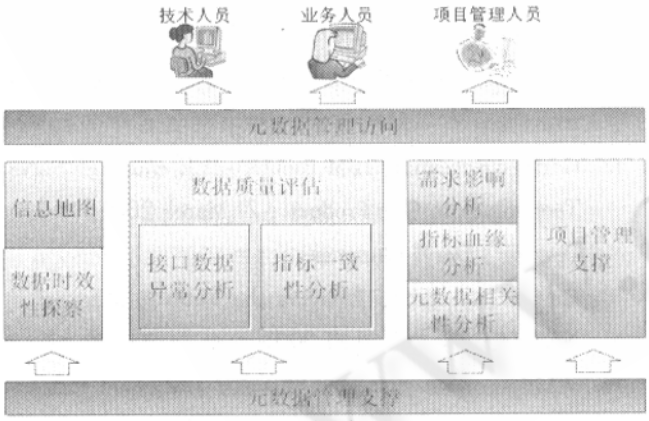

图 4　元数据管理平台功能示意

其中各个层说明如下：

(1) 元数据源层。元数据源层包括经营分析系统的数据源系统，ETL 工具、数据仓库产品、数据集市产品、OLAP 服务器、前端展现工具、数据挖掘工具等。

(2) 元数据获取层。元数据获取层实现元数据源层中各个系统的元数据抽取。

(3) 元数据存储层。元数据存储层实现元数据的存储，元数据按照主题组织。

(4) 元数据管理层。元数据管理层提供 CORBAIDL 接口实现/JMI 接口实现和 XMI 接口实现；实现元数据查询、元数据浏览、元数据访问、元数据分析、元数据导入、元数据导出等基本功能模块。

(5) 元数据访问层。元数据访问层包括元数据管理工具前端、数据仓库系统和中央元数据抽取服务器。这些系统通过元数据管理层访问元数据存储层的元数据。

4.3 应用体系结构(参见图 4)

在元数据管理平台建成之后，其一可以实现对技术元数据的抽取，把相关的字段放到平台上来。在这个平台上，就能清晰地看到这些表或字段之间的关联关系，有一个很清晰的视图。其二还会把业务元数据抽取出来，确定要做哪些应用，就把相关的指标、流程在平台上建立起来。把这些元数据抽取出来后，用户可以通过平台很方便地修改数据仓库中的数据，调整业务中的统计指标等等。其三就是要把技术元数据和业务元数据两种数据对应起来。

从系统维护来看，元数据管理平台使得数据仓库以及业务系统中的各种修改变得省心省力。比如对数据库中表的修改，小的数据仓库模型的修改等等，都可以通过元数据管理平台来实现。同时对数据仓库、OLAP、ETL 等各个层面进行修改。而在以前由于这些工作分散在不同的系统中，耽误时间不说，修改是否准确也不能保证，而业务也会陷于停滞。

从应用分析上看，目前可见的应用主要有三类。

其一，作为即席查询工具做指标的管理，即通过基于元数据的指标管理，掌控各种指标的异常波动情况。

其二，血统分析和影响分析。血统分析指发现某报表中的指标不正常就需要查出问题可能出在哪里。

通过血统图就可以很快找出问题是在源数据系统中，还是在 ODS 层或者是 DW 层中。影响分析则和血统图相反，主要看在修改一个表之后会影响到上游的哪些数据。其三，表重要程度分析和表无关程度分析。主要就是针对现在数据仓库提供的表的数量太多（上万个），通过元数据管理平台就可以列出不同重要程度的表。

5 结束语

元数据管理平台是数据仓库发展的需求，数据质量管理系统实施的核心是元数据管理支撑功能的实施。因此有必要引入元数据的相关功能，在基于元数据管理的基础上建立数据质量管理系统，并通过制定相关管理流程来保证经营分析数据质量的日常管理。

参考文献

1 陈武、袁国忠译，企业数据仓库规划、建立与实现，北京，人民邮电出版社，2000 年 5 月.

2 彭蓉、刘进、何璐璐、刘超，公共仓库元模型：数据仓库集成标准导论企业数据仓库规划、建立与实现，北京，机械工业出版社，2004 年 5 月.

3 W. H. Inmon, Building the Data Warehouse（第二版），王志海等译，北京，机械工业出版社，2000 年 5 月.

4 Jiawei Han, Micheline Kamber, 数据挖掘：概念和技术，北京，机械工业出版社，2001 年 8 月.

5 Alex Berson et al, 贺奇等译，构建面向 CRM 的数据挖掘应用，北京，人民邮电出版社，2001 年 8 月.

6 潇湘工作室译，数据仓库技术指南，北京，人民邮电出版社，2000 年 8 月.

7 SAS Institute, SAS 8.2 OnlineDoc, 2001.

8 Stone, M. et al, Database marketing and customer recruitment, retention and development: what is the technological state of the art? Journal of Database Marketing 1998, 5(4), 303 - 331.

9 Janny C. Hoekstra, Eelkok. R. E. Huizingh, The Lifetime Value Concept in Customer - Based Marketing, Journal of Market Focused Management, 1999 年 3 月份, 257 - 274.

10 P. N. Spring, P. C. Verhoef, J. C. Hoekstra, P. S. H. Leeflang, The Commercial Use of Segmentation and Predictive Modeling Techniques for Database Marketing, 2000, Working Paper, University of Groningen.

11 Michael J. Shaw et al., Knowledge management and data mining for marketing, Decision Support Systems 2001 年 31 期, P127 - 137.

12 J. A. Harding, B. Yu. Information - centered enterprise design supported by a factory data model and data warehousing. Computer In Industry, 1999, Vol. 40, pp. 23 - 36.

13 Keen P G W & Scott - Morton M. Decision Support Systems—An Organizational Perspective. Addison - Wesley, 1978.

14 Shanks. Graeme, Darke. Peta. Understanding corporate data models. Information & Management, 1999, Vol. 35, pp. 19 - 30.

15 Gartner Research, Microsoft Retires COM , 13 July 2000 , Yefim Natis , Daryl Plummer.

16 Gartner Research, Microsoft's Whistler Gets a Name: Windows . NET Server, 26 June 2001 Thomas Bittman.

17 Gartner Research, IBM Blends J2EE and Web Service Technology in One Product , 21 March 2001, Massimo Pezzini , Daryl Plummer.

18 Gartner Research, Weaving the Future: Microsoft's . NET Initiative, 5 April 2001.

19 Whit Andrews , Thomas Bittman , Michael Calvert , Mark Driver , Chris Le Tocq , Daryl Plummer , David Smith.

20 Forrester Report , Putting J2EE to Work, July 2001, Laura Koetzle with Harley Manning, Katharine M. Gardiner, Joyce Tong.