企业数据仓库元数据管理孤岛困境的解决方案探讨

姚晓辉

(中国电信股份有限公司上海研究院 上海 200122)

摘 要 本文首先结合电信企业(中国电信)实际案例,阐述了企业元数据管理的应用现状和困境,分析了元数据管理孤岛困境的成因,提出了电信企业信息系统元数据管理的功能框架,然后结合企业管理流程优化的经验,研究并提出了改善元数据管理的建议以及元数据管理如何融入企业系统建设与运营维护的可行方案。

关键词 元数据;孤岛困境;流程优化;数据质量监控

1 引言

元数据(metadata)最常见的定义是“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。这些描述涉及技术属性特征(例如结构和行为)、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)。早在20世纪末,元数据的概念和相关管理工具就已经出现,但由于当时的数据量还不够大,而元数据本身又包含了太多的内容,以至于并未得到充分利用。

中国电信在企业信息化过程中注重系统建设和维护,很多省公司在数据仓库建设中都考虑了元数据管理工具的部署,但在应用方面却并不如人意。目前元数据管理工具有很多,理论上讲,用户可以用其中任意一种管理其他系统中的数据,比如选择数据仓库系统厂商提供的元数据管理工具来管理其他系统的元数据。但实际应用中的管理效果并不好,一般情况是这些专门工具管理本系统的元数据尚可,一旦跨系统管理,效果就不尽如人意了,很容易形成元数据“信息孤岛”的情形。

2 元数据的分类及应用现状

2.1 元数据的分类

元数据是描述数据仓库内数据的结构和建立方法的数据,可将其按用途的不同分为两类:技术元数据和业务元数据。在数据仓库系统中,元数据可以帮助数据仓库管理员和开发人员非常方便地找到他们所关心的数据。

技术元数据是存储数据仓库系统技术细节的数据,用于开发和管理数据仓库,它主要包括以下信息:

- 数据仓库结构的描述,包括仓库模式、视图、维、层次结构和导出数据的定义以及数据集市的位置和内容;
- 业务系统、数据仓库和数据集市的体系结构和模式;
- 汇总用的算法,包括度量和维定义算法、备份历史、存档历史、信息传输历史、数据获取历史、数据访问以及数据粒度、主题领域、聚集、汇总和预定义的查询与报告访问权限等;
- 由操作环境到数据仓库环境的映射,包括源数据及其内容、数据分割、数据提取、清理、转换规则、数据

刷新规则、安全(用户授权和存取控制)。

业务元数据从业务角度描述了数据仓库中的数据,它提供了介于使用者和实际系统之间的语义层,使得不懂计算机技术的业务人员也能够"读懂"数据仓库中的数据。业务元数据主要包括以下信息:使用者的业务术语所表达的数据模型、对象名和属性名;访问数据的原则和数据的来源;系统所提供的分析方法、公式和报表的信息。具体包括以下信息。

- 企业概念模型:业务元数据所应提供的重要信息,它表示企业数据模型的高层信息、整个企业的业务概念和相互关系。以企业模型为基础,不懂数据库技术和 SQL 语句的业务人员对数据仓库中的数据也能做到心中有数。
- 多维数据模型:企业概念模型的重要组成部分,它告诉业务分析人员在数据集市当中有哪些维、维的类别、数据立方体以及数据集市中的聚合规则。其中,数据立方体表示某主题领域业务事实表和维表的多维组织形式。
- 业务概念模型和物理数据之间的依赖:以上提到的业务元数据只是表示出了数据的业务视图,这些业务视图与实际的数据仓库或数据库、多维数据库中的表、字段、维、层次等之间的对应关系也应该在元数据知识库中有所体现。

2.2 元数据与数据仓库系统

在数据仓库系统中,元数据机制主要支持以下 5 类系统管理功能:

- 描述哪些数据在数据仓库中;
- 定义要进入数据仓库中的数据和从数据仓库中产生的数据;
- 记录根据业务事件发生而随之进行的数据抽取工作的时间安排;
- 记录并检测系统数据一致性的要求和执行情况;
- 衡量数据质量。

与其说数据仓库是软件开发项目,不如说它是系统集成项目,因为它的主要工作是把所需的数据仓库工具集成在一起,完成数据的抽取、转换、加载以及 OLAP(on-line analysis processing,联机分析处理)和数据挖掘等。如图 1 所示,数据仓库的典型结构由操作环境层、数据仓库层和业务层等组成。

其中,第一层(操作环境层)包括整个企业内有关业务的 OLTP(on-line transaction processing,联机事务处理)系统和一些外部数据源;第二层是通过把第一层的相关数据抽取到一个中心区而组成的数据仓库层;第三层是为了完成对业务数据的分析而由各种工具组成的业务层。元数据管理在整个数据仓库系统中起到了承上启下的作用,具体体现在以下几个方面:

- 便于集成;
- 提高了系统的灵活性;
- 保证了数据的质量;
- 帮助用户理解数据的意义。

2.3 电信企业元数据管理的应用现状

目前中国电信使用的元数据工具可提供以下具体功能。

(1)元数据基础功能

元数据模型管理:基于 CWM (common warehouse metamodel,常规数据仓库元数据模型)的定制扩展,用户自定义私有模型。

元数据基础维护:元数据的增、删、改维护和元数据对象的关系维护。

(2)查询检索功能

元数据字典:对结构化存储的元数据信息进行高效便捷的查询检索,多种检索条件可适用于不同用户群体。

(3)应用分析功能

元数据的经典应用:血统分析与影响分析。

元数据的数据扩展分析:元数据差异分析、一致性分析。

(4)导入导出功能

导入功能:用于获取非结构化的元数据信息,是元数据自动获取接口的重要补充。

导出功能:将元数据导出为多种格式的文档,为用户提供了离线的元数据共享和使用方式。

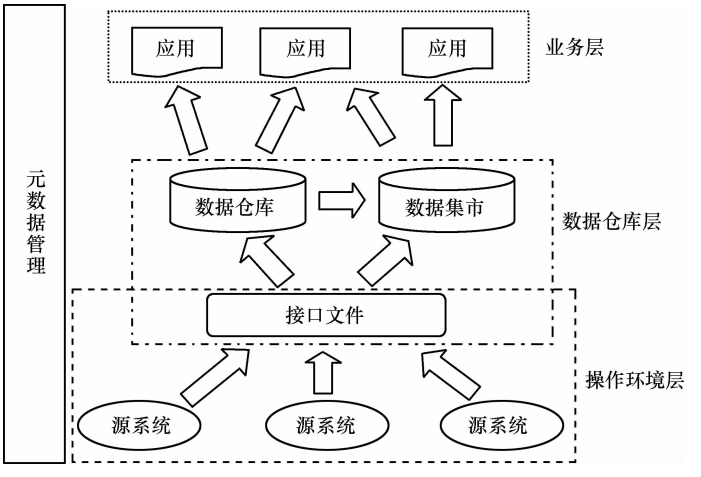

图 1 数据仓库的典型架构

以下列出元数据常用的几个应用场景。

- 应用场景 1:业务人员面对业务指标数据错误、概念模糊的情况时，可以通过元数据工具来了解这个指标的定义、解释以及系统中的数据来源和处理业务规则,以帮助业务人员清晰地把握业务逻辑,明白业务指标的数据错在哪里。
- 应用场景 2：业务人员在平时往往会遇到相同的指标在不同部门理解不一致、相类似指标的差异点、新旧指标的不同定义等需要大量解释工作的问题,可以利用元数据工具查询这些指标的定义、标准、规则及与其他数据的关系等信息，让业务人员更容易地把握这些信息,减少解释的难度。
- 应用场景 3：开发人员在遇到系统升级、变更配置时,往往因为文档不完善或者系统过于复杂,导致对变更引起系统问题准备不足,影响系统正常应用。如果使用元数据工具就能很好地解决这个问题，对升级系统所产生的影响做全面评估，制定合理的变更计划并在系统中实施,以保证系统平稳运行。

总结上述应用场景，可以发现元数据应用的关键作用体现在以下关键场景中：

- 岗位变动,新人急需了解系统构架、数据现状；
- 业务咨询,需要了解数据来源、口径及对数据做了哪些处理；
- 数据出现异常,需要快速定位,维护人员无法逐个查找 SQL 文档；
- 日常业务人员、维护人员需要直观的图形化元数据管理工具；
- 数据质量管控的高要求。

但是在使用元数据工具时，笔者发现工具并不能解决所有的问题,在管理跨系统的元数据时,往往因更新不及时而导致无法取得元数据的信息，元数据管理系统大多数时候成为了“信息孤岛”,应用效果大打折扣。

3 元数据管理孤岛困境的成因分析

元数据管理问题的成因很容易总结，即元数据管理相对独立于系统,未嵌入流程管理,一旦维护不及时,元数据管理系统就容易成为信息孤岛,失去作用。

目前，中国电信大部分省公司的元数据管理都没有使用专业工具,而在购买了专业工具的省公司中,元数据管理工具的使用也不令人满意。元数据管理工具的部署在初期往往需要做大量底层配置工作,在目标系统布设监控点,一旦元数据信息发生变更，就需更新元数据管理系统的数据库。但是目前元数据管理系统往往都采用被动要数的方式,不主动抓取数据,完全依靠目标系统,受制于目标系统能否及时提供元数据信息,因此在初期一次性配置完成后,后期完善维护的工作量将相当大,如果没有流程和制度来保证,元数据管理系统在目标系统的若干次调整之后，必然跟不上变化,沦为“信息孤岛”,失去作用。

4 元数据管理流程优化建议

中国电信经过多年 EDA(enterprise data architecture,企业数据架构)建设,各省公司都积累了一些经验和教训,元数据管理的关键还在于完善规范的管理流程和先进易用的技术工具。

笔者根据实践经验提出了元数据管理功能框架,如图 2 所示。

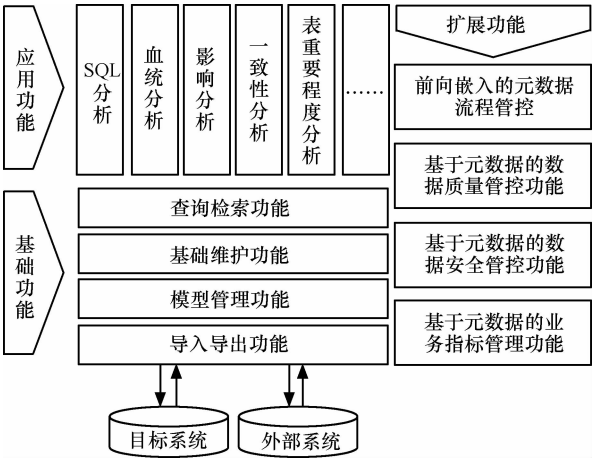

图 2　元数据管理功能框架

4.1 嵌入流程管理

很显然，解决元数据管理困境的方法在于规范管理流程,变后置的元数据更新方式为前置的更新方式,把元数据管理纳入企业信息数据核心管理流程，将其和数据质量管控结合起来,提高数据管理的效率。

以企业指标管理为例，可通过元数据系统的管理流程功能支撑实现对指标的管理。如图 3 所示,指标管理流程主要涉及 4 个环节:指标定义与变更环节、指标审核环节、指标维护与实现环节、指标发布环节。把元数据的变更、审核、实现等环节嵌入到指标变更环节中，使得元数据管理系统与目标系统的配置保证是实时一致的。

4.2 结合数据质量管理

一方面,可以通过元数据来提升数据质量的管理效率,

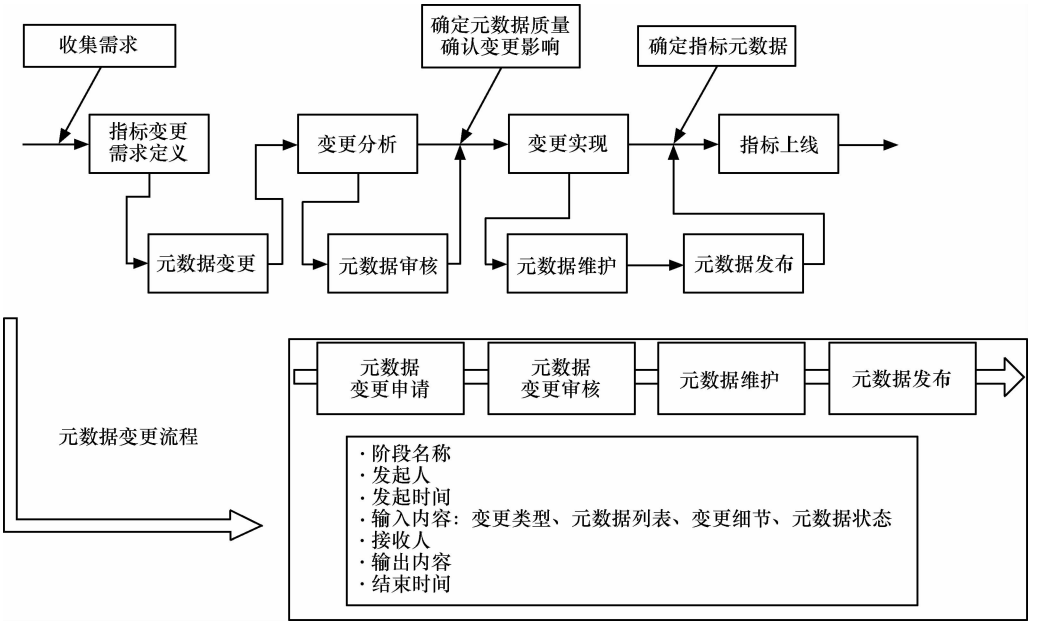

图 3　前向嵌入的元数据管控流程

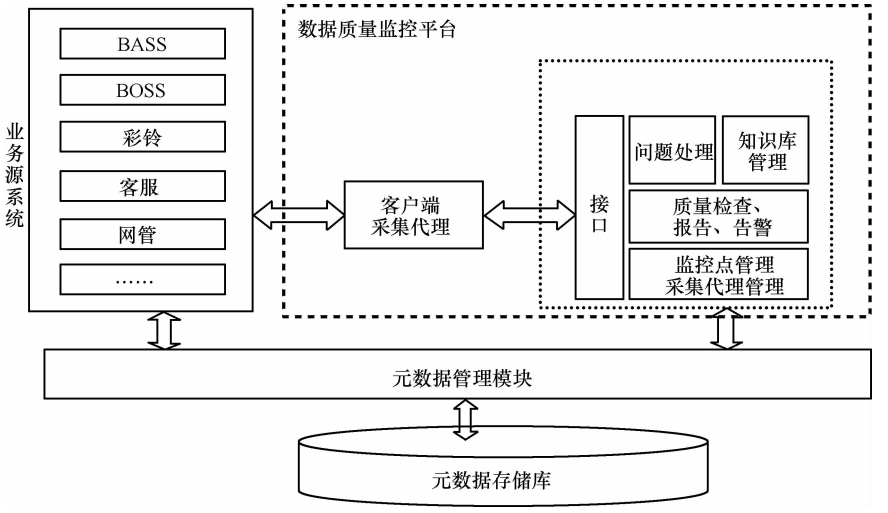

图 4　基于元数据的数据质量监控系统框架

另一方面，也可以把元数据管理嵌入到数据质量管控流程中,使元数据管理得到充分的重视。如图 4 所示。

应用 1:利用元数据分析监控点的部署情况(从各个不同视角的数据处理流程中规划哪些对象应该被监控，分析哪些内容被监控),如图 5 所示。

应用 2:通过不同主线组织数据处理流程的视角,在数据处理流程中分析各个环节的数据质量问题发生、处理情况。

- 通过多个维度对经营分析系统的整体数据质量环境进行数据质量视角划分。
- 确保所有视角完整覆盖整个经营分析系统的数据质量环境,每个视角有明确的数据质量监控主题。
- 根据视角的划分进行监控部署的总体布局，建立完整的监控点树。

应用 3:在数据质量问题的分析处理中,通过元数据的血统分析功能分析问题产生的根源,通过元数据的影响分析功能分析该问题对后面哪些环节有影响,以确定下一步处理策略。

从上述具体应用可以看出,借助于元数据,数据质量管控工作可以大大提高效率，为数据质量监控后的维护工作带来了极大的帮助。

4.3　加强业务元数据与技术元数据关系的建设

建立起业务指标到报表的关系，通过系统动态配置业务指标与系统维度代码之间的映射关系，动态管理业务指

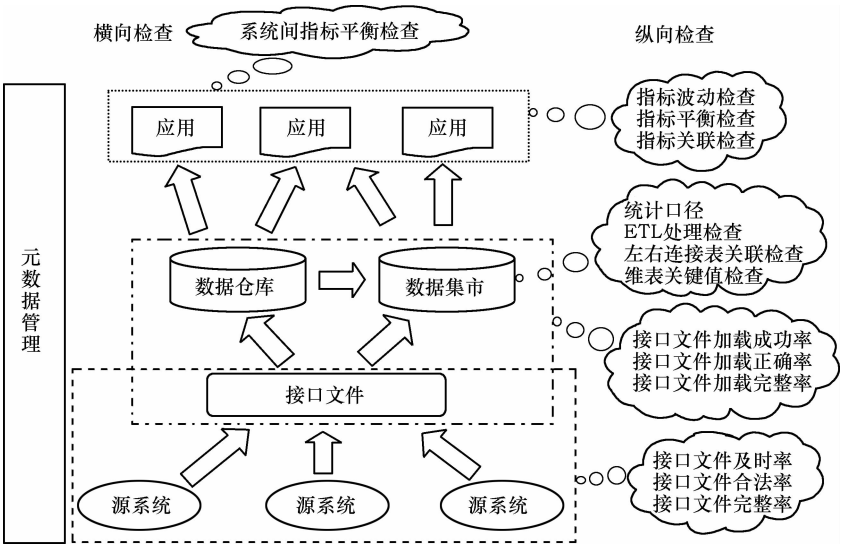

图 5　各个环节监控的主要问题

标的业务逻辑与规则。

加强应用和功能之间的关系,体现业务核心指标、KPI、功能服务之间的关联、依赖关系,提高重用性,减少重复建设。

加强业务和人员之间的关系，体现业务有哪些人员与之相关联,在业务变化时有针对性地进行通知。

在数据、功能、维护人员等各方面加强元数据管理功能,使元数据更易管理和使用,以提高元数据管理的水平。

5　结束语

元数据管理是企业信息系统管理的重要领域，是提升企业系统维护和管理能力的基础，是提高企业信息数据使用效率的催化剂,是企业数据质量的重要保证。元数据特别是业务元数据是企业的重要财富，目前人们已经慢慢意识到管理好、用好这些数据是企业提升内部竞争力的重要手段,而在推进过程中遇到的元数据“信息孤岛”的问题其实是流程管理制度上的问题,工具再强大,管理跟不上也是不行的。所以,如何建立和优化企业的信息数据管理流程,把元数据管理真正嵌入到企业的核心流程中，保证元数据的“活性”,避免孤岛的形成,成为了需要面对的主要问题,值得深入探讨。

参考文献

1　吴显义. 我国元数据研究现状分析. 情报科学,2004(1)

2　张英朝. 数据仓库元数据管理研究. 计算机工程,2003,29(1)

Enterprise Data Warehouse Metadata Management Solution for Islanding Dilemma

Yao Xiaohui

(Shanghai Research Institute of China Telecom Co., Ltd., Shanghai 200122, China)

Abstract　This article first combined telecom company (China Telecom) real case, explained enterprise metadata management application status and predicament, analyzed the metadata management in the cause of islanding dilemma,proposed a telecommunications enterprise information system metadata management feature framework, and then combined business management experience in business process optimization,study and improve the proposal of the metadata management, and metadata management how to integrate into the construction and maintenance of enterprise information system.

Key words　metadata, islanding dilemma, business process optimization, data quality monitor　(收稿日期:2009-07-15)