数据仓库元数据管理

余友波

数据仓库之路原创资料

http://www.dwway.com

1.1.1 第一章 元数据概论

企业的计算机系统每年会产生很多数据，很多企业面临着这样的困境，难以有效的管理大量的、繁杂的、不一致的数据，并方便地访问、利用这些数据进行辅助决策。

建立数据仓库提供一个方法，把数据转化为有用的、可信赖的信息，支持商业决策。建立数据仓库一个重要的工作是元数据管理。元数据（Metadata）就是数据的数据，用于建立、管理、维护和使用数据仓库。。元数据管理是企业级数据仓库中的关键组件，贯穿于建立数据仓库的整个过程。

元数据使得用户可以掌握数据的历史情况，如数据从哪里来？流通时间有多长？更新频率是多大？数据元素的含义是什么？对它已经进行了哪些计算、转换和筛选等等。在需求不确定情况下，在瞬间万变的商业环境下，元数据可以更好的支持需求的变化，降低项目风险。

通常把元数据分为技术元数据（Technical Metadata）和业务元数据（Business Metadata）。技术元数据是描述关于数据仓库技术细节的数据，这些元数据应用于开发、管理和维护数据仓库；业务元数据从商业和业务的角度描述数据仓库的数据，提供了良好的语义层定义，业务元数据使业务人员能够更好的理解数据仓库分析出来的数据。

元数据贯彻于建立数据仓库的整个过程，不只是 ETL 过程需要元数据的支持。

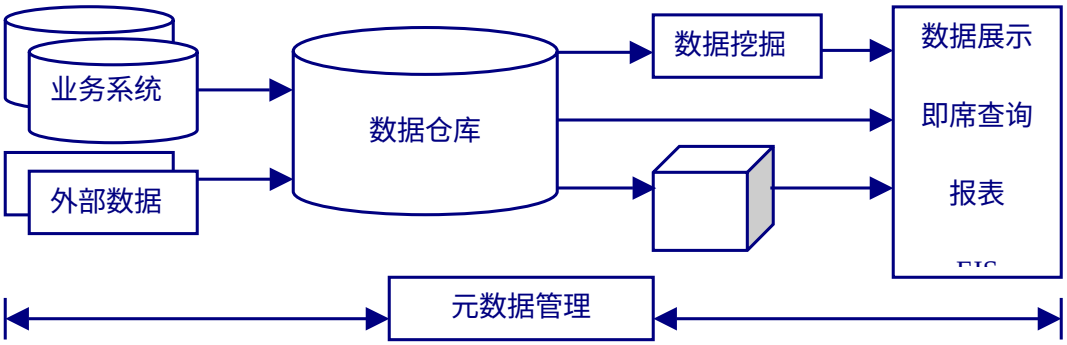

图 1 元数据的应用

在使用元数据的同时，随着数据仓库市场的发展，业界出现许多数据仓库管理和分析的工具，各种工具使用不同的元数据标准来表示和处理，不同系统之间的迁移、数据交换变得困难。于是，我们希望用一种单一的元数据标准，使得各种组织的元数据具有单一的元模型（MetaModel），因此，需要建立一种标准使得不同的数据仓库和商业智能系统之间可以相互交换元数据。

1.1.2 第二章 元数据标准

1.1.2.1 一、元数据标准 CWM

OMG于2001年颁布元数据标准CWM 1.0（Common Warehouse Metamodel Version 1.0）。CWM定义一个描述数据源、数据目的、转换、分析的元数据框架，以及定义建立和管理数据仓库的过程和操作，提供使用信息的继承。

目前宣布支持CWM的厂商包括：IBM、Oracle、Hyperion、Dimension EDI、Genesis IONA、HP、NCR和Unisys等。

CWM基于3个工业标准：

- UML - Unified Modeling Language，OMG建模标准；
- MOF - Meta Object Facility，OMG建立元模型和模型库的标准，提供在异构环境下的数据交换的接口；
- XMI - XML Metadata Interchange，OMG元数据交换标准。

UML在CWM中得到充分的应用，担任3个不同的角色：

1），UML用来做为与MOF对应的meta-metamodel。UML相当于MOF Model,，UML Notation和OCL(Object Constraint Language)，被用来做为建模语言、图形符号、约束语言，定义和描述CWM。

2），UML用来创建元模型。UML，特别是Object Model 包描述的子集，用来从其它元模型继承等级和关联以建立CWM。

3），UML做为面向对象元模型（object-oriented metamodel）。UML被用来描述面向对象的数据。

CWM元模型包括大量的子元模型（sub-Metamodel），这些子元模型描述了建立数据仓库和商业智能的各个主要部分的通用数据仓库元数据。

主要包括：

1）、数据资源：包括各个元模型，描述了面向对象数据、关系数据库、记录、多维和XML等数据。

2）、数据分析：包括描述数据转换、OLAP、数据挖掘、信息展现、商业术语等的元模型。

3）、数据仓库管理：这包括数据仓库过程以及数据仓库操作结果的元模型。

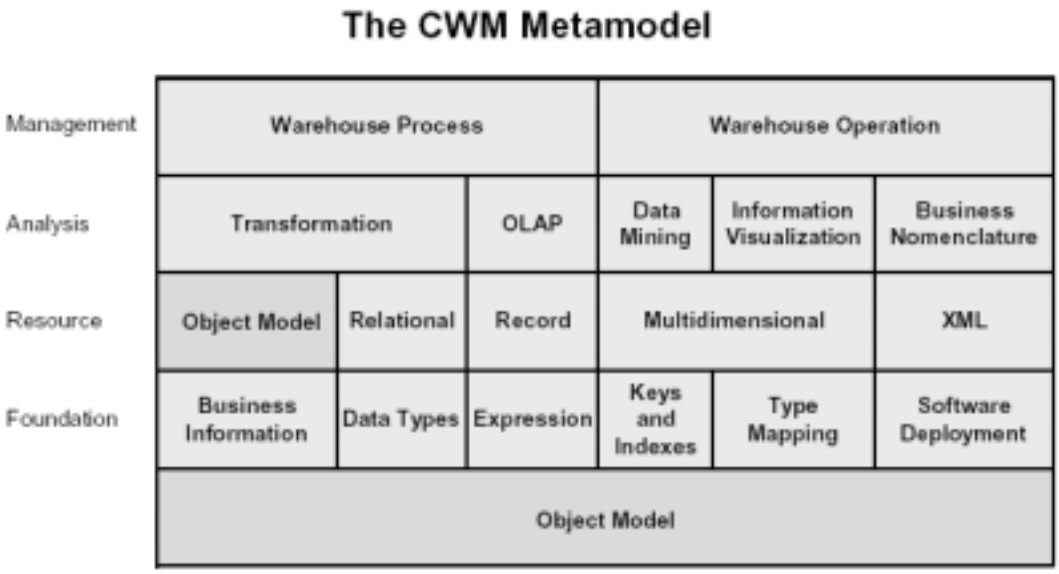

图2 CWM元模型架构图

CWM元模型设计的目的是最大化的重用对象模型Object Model (UML的子集)，尽可能的共享通用的模型构建。最典型的是，CWM重用/依赖对象模型来描述面向对象的数据资源；另外，其它类型的数据资源的主要Metamodel元素，在对象模型中都有相同的模型元素与之相对应。

1.1.2.2 二、使用 CWM

1、CWM 的目标使用者

CWM 标准包括了技术元数据和业务元数据的定义，涉及数据仓库生命周期的所有阶段，所以不只是实施工程师和实施顾问使用 CWM，最终用户也会受益于CWM。

CWM 的目标使用者包括 6 类人员：

1，数据仓库平台和工具供应商；

2，专业服务咨询商；

3，数据仓库开发者；

4，数据仓库管理员；

5，最终用户；

6，信息技术主管（CIO）。

2、基于 CWM 的数据仓库

CWM 的目标使用者将会参与到开发和使用基于 CWM 的数据仓库的过程中；但并不是所有的角色需要参与整个过程，而是参与到下面列举的的 4 个阶段中的一个或多个：

1）、Establishment。实现和配置CWM，包括建立一个通用资料库。

2）、Build。使用CWM定义一个基线数据仓库配置（建立数据源和目的的交换路径）。

3）、Operation。操作和使用基于CWM的数据仓库。

4）、Maintenance。维护使用了CWM定义的数据仓库的配置。

1.1.2.3 三、CWM 标准组织结构

CWM元模型使用包（package）和包等级结构来控制复杂性、提高理解性、支持重用。模型元素包括下面的包：

1， 对象模型包

对象模型包是构建和描述其它 CWM 包的元模型类的基础。

- 核心包。包括CWM核心对象模型的类和关联，被其它CWM包使用。
- 行为包。包括用来描述CWM对象的行为的类和关联。

- 关系包。包括用来描述各个CWM对象之间关系的类和关联。
- 实例包。包括用来描述CWM实例的类和关联。

2、基础包

基础包是表示 CWM 概念和架构的模型元素。

- 商业信息包。包括用来描述关于模型元素的商业信息的类和关联。
- 数据类型包。包括用来描述创建模型需要的特定数据类型构建的类和关联。
- 表达式包。包括用来描述表达树（expression trees）的类和关联。
- 关键字和索引包。包括用来描述主键和索引的类和关联。
- 软件部署包。包括用来描述软件在数据仓库中如何部署和配置的类和关联。
- 类型映射包。包括用来描述两个系统之间数据类型映射关系的类和关联。

3、资源包

资源包是用来描述数据资源和记录的信息。

- 关系包。包括用来描述关系型数据的元数据的类和关联。
- 记录包。包括用来描述记录型数据的元数据的类和关联。
- 多维包。包括用来描述多维型数据的元数据的类和关联。
- XML包。包括用来描述XML数据的元数据的类和关联。

4、分析包

分析包定义了如何对信息进行加工和处理，以及信息展示。

- 转换包。包括用来描述数据转换工具的元数据的类和关联。
- OLAP包。包括用来描述OLAP工具的元数据的类和关联。
- Data Mining包。包括用来描述数据挖掘工具的元数据的类和关联。

- 信息展示包。包括用来描述信息展示工具的元数据的类和关联。
- 商业术语包。包括用来描述商业分类学和术语表的元数据的类和关联。

5、管理包

管理包用于数据仓库管理和维护。

- 仓库过程包。包括用来描述数据仓库过程的元数据的类和关联。
- 仓库操作。包括用来描述数据仓库操作和查询结果的元数据的类和关联。

1.1.3 第三章 建立元数据库

元数据库是用于存储元数据的地方，元数据库最好选用主流的关系数据库管理系统，支持 CWM 标准。一个元数据库还包含那些用于操作和查询元数据的机制；建立元数据库的主要好处是提供了统一的关键数据结构和业务规则，易于将企业内部的多个数据集市有机的结合起来，特别是，现在一些客户倾向建立多个数据集市，而不是一个庞大无比的数据仓库。

可以考虑在建立数据仓库（或数据集市）之前，先建立一个用于描述数据的、用于应用集成的元数据库，做好数据仓库实施的初期支持工作，对后续开发和维护有很大的帮助。

在拥有不同厂商、不同功能和不同元数据库的环境下，要实现两种产品之间的元数据同步是非常富有挑战性的工作。因为必须从一种产品中获得足够详细的元数据，将其映射到另一种产品中，再指出两者意义或编码的差别；通常系统有数百、数千个元数据，必须对每个元数据重复这一过程。

在整个数据仓库环境中，元数据管理工具可以从各个数据仓库组件中收集元数据，存储到元数据库中，然后向业务用户传递和展示正确的信息。采集、集成和描述元数据可以扩展到十分广泛的范围，可以在设计和建模的过程中，可以在数据转换、清洗和过滤的过程中，也可以在数据移植的过程中；可以从数据库/数据存储软件，和前端展示工具中得到元数据。

元数据库为整个企业的宝贵信息提供了详细的记录，保存数据存储位置和商业含义、生成和维护数据的主体、数据驱动的应用处理、与其它数据的关系以及数据的转换过程等。元数据库保证了数据仓库数据的一致性和准确性，为企业进行数据

质量管理提供数据依据。

另外，元数据库还支持强大的查询和报表生成工具，用户使用报表工具可以查询元数据库，从元数据库获得重要的决策支持信息。