基于目录的数据管理与发布模型的研究与应用

沈志宏，吴开超
中国科学院计算机网络信息中心，北京 100190

摘 要：在长期的科研活动中，科研人员积累了大量的结构化数据，大部分数据会采取关系型数据库管理系统作为存储，如何对这些数据进行简单、直观的管理，并开发应用程序对这些数据进行发布，已成为一个比较棘手的问题。本文基于元数据目录，提出了一套相对通用的数据管理与发布模型，并基于该模型开发VisualDB软件，为科研人员提供了可视化的数据管理、发布界面，使建库人员从繁琐的数据库管理、开发工作中解放出来，目前该工具在科学数据库项目及其它数据库项目中得到了广泛的应用。

关键词：科研数据 元数据 目录 关系型数据库 数据管理 数据发布

Research on Catalog based Data Management and Publish Model

Shen Zhihong, Wu Kaichao
Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190 China

Abstract: In the long term scientific activities, scientists have collected lots of structural data and stored them in rational database management system. How to manage these data and publish them in a easy way is a problem. This paper introduces a data management and publish model based on metadata catalog, and a management tool VisualDB, which provides visual user interface for data owners. By now, VisualDB has been used in some projects; it is expected to be more widely used in more scientific data projects.

Keywords: scientific data, metadata, catalog, rational database management system, data management, data publish

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1.引　言

在长期的科研活动中，科研人员积累了大量高度结构化的数据，除了以文件格式实现存储，大部分科研数据会采取关系型的数据库管理系统（RDBMS）作为存储，这包括ORACLE、SQLServer、MySQL、Access等。

基于关系型数据库进行数据的管理，需要直观、方便的客户端软件来提供支持。一般RDBMS的产品都会提供相应的数据库管理软件，如：SQL Server的SQL Server Management Studio，Oracle的SQL Enterprise Manager，等等。然而，这些工具对用户的数据库专业知识要求都比较高，界面比较复杂，同时缺乏对数据编辑以及关联关系处理的友好界面。在访问控制方面也比较复杂，一般都需要专门的数据库管理员进行操作。此外，部分文件型数据库，如：Microsoft Access，其对分布式管理的功能有限，这就给多人协作管理数据库带来一定的困难。

另一方面，为了及时共享科学数据，科研工作者往往需要投入精力完成数据发布程序的开发。此外，随着科研工作的不断深入，数据的结构很有可能发生更改和补充，由此带来的应用系统的升级和维护也是一个比较重要的问题。

为了解决这些问题，我们通过多年的研究，提出了一套基于目录的数据管理与发布模型，并成功基于此模型开发了软件工具VisualDB 1.0，使用该工具可以使建库人员从繁琐的数据库管理、开发工作中解放出来。目前VisualDB已成功应用到包括物理化学、医药卫生、人文、政务、文献、财税等领域的十多个项目中。

2.元数据与目录

(1)元数据

元数据是关于数据的数据（data about data），它用于描述数据的内容（what）、覆盖范围（where/when）、质量、管理方式、数据的所有者（who）、数据的提供方式（how）等信息，是数据与数据用户之间的桥梁。元数据可以为各种形态的信息资源提供规范、普遍的描述方法和检索工具，为分布的、由多种资源组成的信息体系提供整合的工具与纽带。

元数据一般可分为描述性元数据（Intellectual Metadata）、结构性元数据（Structural Metadata）、存取控制性元数据（Access Control Metadata）、评价性元数据（Critical Metadata）等等。描述性元数据通常用来描述、发现和鉴别数字化信息对象，它主要描述信息资源的主题和内容特征。结构性元数据描述数字还信息资源的内部结构，如数据集的库、表、字段的特征。存取控制性元数据用来描述数字化信息资源能够被利用的基本条件和期限，以及这些资源的知识产权特征和使用权限。评价性元数据描述和管理数据在信息评价体系中的位置。

可以说，元数据标准和技术是实现数据管理与共享的主要手段之一。通过元数据可以达到数据标准化以及数据共享、交换和整合。在数据管理与发布的模型中，包含有各种类型的元数据，并最终已一定的形式构成元数据目录。

(2)数据模型

数据模型用来描述一组数据的概念和定义，包括概念数据模型（Conceptual Data Model）、逻辑数据模型（Logical Data Model）和物理数据模型（Physical Data Model）。概念数据模型面向数据库用户的实现世界的数据模型，主要用来描述世界的概念化结构，它使数据库的设计人员在设计的初始阶段，摆脱计算机系统及DBMS的具体技术问题，集中精力分析数据以及数据之间的联系等，与具体的DBMS无关。概念数据模型必须换成逻辑数据模型，才能在DBMS中实现。逻辑数据模型即用户从数据库所看到的数据模型，是具体的DBMS所支持的数据模型，如网状数据模型(Network Data Model)、层次数据模型(Hierarchical Data Model)等等。此模型既要面向用户，又要面向系统，主要用于数据库管理系统（DBMS）的实现。物理数据模型描述数据在储存介质上的组织结构的数据模型，它不但与具体的DBMS有关，而且还与操作系统和硬件有关。每一种逻辑数据模型在实现时都有起对应的物理数据模型。DBMS为了保证其独立性与可移植性，大部分物理数据模型的实现工作由系统自动完成，而设计者只设计索引、聚集等特殊结构。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

在数据管理与发布的模型中，针对不同层次的数据模型，系统会提供不同层次的管理工具和访问接口。VisualDB工具则主要侧重于逻辑模型及概念模型的管理，最大限度的屏蔽了物理模型，让建库人员轻松简便的完成数据库的建设。

图1：基于目录的数据管理与发布模型

(3)元数据目录

元数据是描述数据的数据，模型系统中的所有元数据构成了元数据目录，它采用一种统一的结构来描述各种元数据。元数据目录为用户身份认证、数据定位、访问控制、数据复制等提供了支持。

根据元数据来源和含义的不同，元数据目录可以分成数据集注册目录、数据集结构目录、应用风格目录以及访问控制目录。数据集结构目录主要存储数据集元数据（Dataset）、实体元数据（Entity）、属性元数据（Attribute），应用风格目录侧重于应用框架、显示风格等的描述，访问控制目录则通过维护一份用户授权关系，实现对共享数据的保护。数据集注册目录存储数据集的基本信息以及位置索引信息，显而易见，基于数据集注册目录，最终用户可以很方便的发现并定位到指定的数据集。

3.系统总体结构

该模型的总体结构可以分解为四层，主要包括数据层（data layer）、核心层（business layer）、服务层（service layer）及表现层（presentation layer），如图1所示。

数据层通过关系型数据库管理系统（RDBMS）和文件系统（FS）完成数据的存储与管理。在物理数据之上，采用ORM（Object-Relational Mapping，对象—关系映射）以及数据适配器（adapter）等技术，屏蔽数据库版本之间的差异，完成对分布式数据的访问。并采取DBCP（Database Connection Pool，数据库连接池）、cache（缓存技术）等手段，提高数据的访问性能。同时采用文件系统，完成对科学数据中CLOB(Character Large Object)和BLOB(Binary Large Object)等大字段的存储和管理。

核心层实现数据访问的功能，主要包括：访问控制框架（Access Controller）、数据集成引擎（Data Integration Engine）、数据统计分析引擎（Data Statistic Engine）、元数据访问中间件（Metadata Accessor）、目录访问中间件（Catalog Accessor）以及数据访问中间件（Data Accessor）。通过采用基于约束的访问控制框架（Constraint Based Access Controller），完成RBAC（Role Based Access Control）以及细粒度的权限判定，从而达到对数据的动态保护，同时采用（SSOSingle Sign On，单点登录）等技术完成对用户的认证管理。数据集成引擎通过数据集映射模型，完成对分布式、异构异质数据的集成，数据统计分析引擎通过在线记帐信息，完成对数据量以及数据访问量的统计，并采用JavaMail、监听器等技术完成统计结果的显示和异常通知。元数据访问中间件完成对元数据的管理，目录访问中间件完成对元数据目录的管理，数据访问中间件针对逻辑模型，完成对各种数据的操作。

服务层通过中间件的协调调用，为本地应用和分布式应用提供可用的服务接口，这些服务主要包括目录服务和数据服务。

表现层分别基于Web环境，J2EE/.NET、基于Windows环境、基于PDA环境，Windows Mobile等开发表现层应用，目前我们已开发J2EE版本和PDA

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

版本。

元数据目录作为数据管理与发布模型的核心，通过采用工作流引擎（WorkFlow Engine）等技术，针对数据的全生命周期提供了全方位的目录支持和数据访问支持，这些模块包括数据采集、数据管理、数据发布、数据维护、数据集成与数据统计，如图2所示。

4.VisualDB 1.0

VisualDB提供了基于目录的数据管理与发布模型的一个实现，目前的发布版本为1.0。VisualDB基于Web可视化界面，为建库人员提供了关系型数据库的数据录入、更新、发布以及安全控制的功能。建库人员通过简单的配置和选择，无需编写代码，即可完成对数据库的内容发布，提供数据的检索和浏览服务。同时，VisualDB为数据应用开发人员提供可配置的应用模块，为二次开发提供了数据访问接口。利用VisualDB管理的数据集，可以自动成为分布式数据集中的一个数据节点，实现VisualDB管理的数据集之间的自由访问和共享。

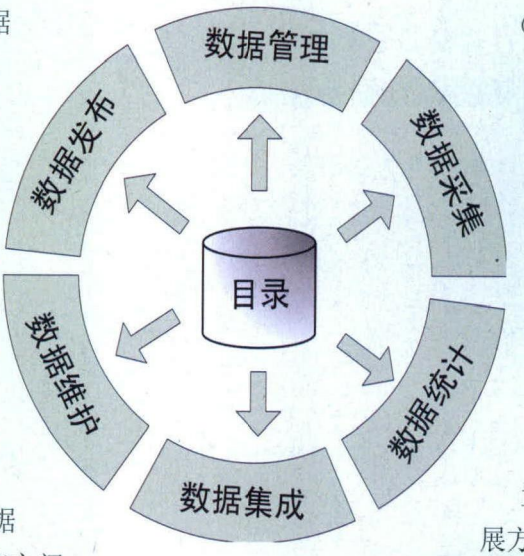

图2：使用目录完成数据的管理与发布

VisualDB 1.0的功能特性主要包括：

■ **完全可视化，高度可配置性，实现零开发。**VisualDB通过完全web化的“目录配置”模块，完成对数据集的配置。用户基本可以脱离后台数据库系统体验数据库的建设。用户无需编写一行代码，只需要通过简单的配置，即可产生数据的管理界面和发布界面，真正做到“零开发”。从而将数据库应用开发平民化，大大节约了传统的数据应用开发带来的开发成本和维护成本。

■ **提供强有力的访问控制，有效保护用户数据。**VisualDB通过其中的“安全中心”模块，将用户进行分级，不同的用户对不同的数据表的查看、修改等权限，都由系统管理员进行分配，从而达到对数据最大限度的保护。此外，访问控制模型具有良好的可扩展性，用户可以通过软件升级或者二次开发，对访问控制的功能进行增强。

■ **普适性。**VisualDB基于流行的B/S结构开发，最终用户只需要IE浏览器即可进行数据的管理和访问。VisualDB支持流行的RDBMS（关系型数据库），并有效屏蔽底层数据库的异构性。支持的RDBMS包括：Oracle、SQLServer、MySQL、Access，以及开源数据库HSQL、SQLite等，此外，VisualDB还支持ODBC数据源。

■ **面向逻辑模型建模，提供丰富的、可扩展的数据类型。**与一般的数据库管理软件不同，VisualDB直接面向逻辑模型（而非面向物理模型），提供丰富的数据类型：主要包括文件、字典、日期、链接、图片、音频、视频、化学结构式、化学分子式、HTML等类型，VisualDB将在后续版本中支持更多的应用级数据类型。

■ **高度灵活的可本地化特性。**VisualDB在提供可配置特性的同时，还提供了一种扩展方案，允许用户自己对应用的界面和行为进行自定义。VisualDB会自动的帮助用户做好数据的发布，但是用户可以通过定制化手段将应用的功能和界面开发的更加符合专业领域的主题风格，如图3所示。

VisualDB 1.0经历了5年多来的开发，实现了基于目录的数据管理与发布模型，目前已在多个项目中得到广泛应用，重点包括纳米科技基础数据库（http://www.nano.csdb.cn/）、中草药数据库（http://www.medicine.csdb.cn）、化学物质毒性数据库（http://www.toxic.csdb.cn/）、中国科学院高级专家数据库（http://www.experts.csdb.cn/）、西北人文数据库、Brainbank数据库（http://dc.brainbank.cn）等。目前VisualDB正在制定新的版本更新计划，新的版本将进一步提高系统的可配置性、可扩展性及可适配性，基于目录的数据建模

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

图3：基于VisualDB本地化的应用系统

过程将变得更加科学、灵活、方便。同时，VisualDB 1.x将增强系统的类型库（type library），除常见的文本、数值、视频、附件等格式，支持更多的科学数据类型，为科学数据的著录和发布提供更好的界面支持。最后，在二次开发接口上VisualDB 1.x需要进一步的优化，让应用开发人员基于领域模型能够更快捷的。

参考文献

[1] 中国科学院计算机网络信息中心科学数据库中心. 中国科学院科学数据库核心元数据标准（版本号：1.1）. 2003年8月.

[2] 沈志宏，王龙潇. 目录型元数据在数据访问系统中的应用. 科学数据库与信息技术论文集，第七集，2004年.

作者信息

沈志宏

中国科学院计算机网络信息中心，科学数据中心，高级工程师，主要研究领域为数据库应用。

吴开超

中国科学院计算机网络信息中心，科学数据中心，高级工程师，主要研究领域为数据库应用。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net