

XML 技术在基于 PowerMart 的项目开发中的应用

吴 军 韩 斌 刘 刚¹

(江苏科技大学计算机科学与工程学院, 镇江 212003; 广州海捷计算机科技有限公司¹, 广州 510665)

摘 要 PowerMart 是一款性能强大的 ETL 工具,但在基于 PowerMart 的项目开发过程中,其界面操作繁琐,开发耗时长。本文通过对 powermart. dtd 的分析,提出了一种基于 XML 技术的开发方式,针对具有类似 ETL 映射或 Workflow 的作业,设计出统一的 XML 模版,通过修改 XML 模版来生成新的 Workflow 作业。在广东发展银行 EDS 项目开发过程中证明了这种方法可以显著提高开发效率。

关键词 PowerMart 文档类型定义 XML 技术 数据抽取 数据转换 数据加载

中图法分类号 TP302. 1; **文献标志码** A

XML 是由 W3C 制定的新标准标记语言,其最大特点是将信息的描述与信息处理分开,使得数据具有自我描述的能力。PowerMart 是一种高性能的 ETL 工具,PowerMart. dtd 是 PowerMart 元模型的一种表示形式,整个 ETL 规则(XML 文档)的语法正是由此定义的,通过它可以剖析 PowerMart 元模型的组织结构。掌握 PowerMart. dtd,可以有效地提高开发效率。

1 XML 技术

XML(extensible markup language)是 W3C 组织于 1998 年 2 月制定的一种通用语言规范,是 SGML 的简化子集,它以一种开放的自我描述方式定义数据结构,在描述数据内容的同时能突出对结构的描述,从而体现出数据之间的关系。XML 是一种元标记语言^[1,2],使用者可按需要创建新的标记,XML 的可扩展性就在于此。带标记的元素是 XML 文档的构造块(如图 1),这种元素可以有若干个属性,并可

以包含零个或多个子元素。这些元素可以是文本数据,也可以是带标记的元素。

```
<? xml version = "1.0" encoding = "gb2312" ?>
<! DOCTYPE POWERMART SYSTEM "powermart. dtd" >
<POWERMART CREATION_DATE = "09/12/2009 09:46:08" RE-
POSTORY_VERSION = "133.41" >
  <REPOSITORY NAME = "REP_NAME" VERSION = "133"
CODEPAGE = "MS936"
DATABASETYPE = "DB2" >
    <FOLDER NAME = "FLD_NAME" GROUP = "Administrators"
OWNER = "Administrator" >
      <CONFIG NAME = "default_session_config" ISDEFAULT = "
YES" >
        </CONFIG>
      <SESSION NAME = "s_M_MAME" ISVALID = "YES" MAP-
PINGNAME = "M_MAME"
MAPPINGVERSION = "1.0.0" >
        </SESSION>
      <WORKFLOW NAME = "WF_NAME" ISVALID = "YES"
SERVERNAME = "SER_NAME"
SUSPEND_ON_ERROR = "NO" REUSABLE_SCHEDULER = "
NO" >
        </WORKFLOW>
      </FOLDER>
    </REPOSITORY>
  </POWERMART>
```

图 1 XML 文档示例

2009 年 9 月 15 日收到

第一作者简介:吴 军(1984—),男,汉族,武汉人,江苏科技大学硕士。E-mail: whwujun@hotmail.com。

XML 文档可以在它的文档类型声明里声明某个 DTD (Document Type Definition, 文档类型定义)。如上述 XML 文档的 DTD 由文件 powermart.dtd 定义^[3] (如图 2)。DTD 是关于 XML 文档中出现的标记和元素结构语法约束。它可以用来验证一个 XML 文档。DTD 是一系列关于元素类型 (Element Type), 属性 (Attributes), 实体 (Entities) 和符号 (Notations) 的定义。它定义了文档所需要的标记, 比如可在文档里使用的元素类型, 以及这些元素之间可能的联系。

```
<! -- Informatica Repository DTD Grammar - Version 6.1 -- >
<! ELEMENT POWERMART(REPOSITORY * )>
<! ATTLIST POWERMART CREATION DATE CDATA #IMPLIED
REPOSITORY NERSION CDATA #IMPLIED>
<! ELEMENT REPOSITORY ( FOLDER * )>
<! ATTLIST REPOSITORY NAME CDATA #REQUIRED
DATABASETYPE CDATA #IMPLIED
CODEPAGE CDATA #IMPLIED
VERSION CDATA #IMPLIED
PARENTREPOSITORY CDATA #IMPLIED>
<! ELEMENT FOLDER ( FOLDERVERDION * , CONFIG * ,
SCHEDULER * ,
TASK * ,SESSION * ,WORKLET * ,WORKFLOW * )>
<! ATTLIST FOLDER NAME CDATA #REQUIRED DESCRIP-
TION CDATA #IMPLIED
SHARED (SHARED|NOTSHARED) #IMPLIED OWNER CDATA
#IMPLIED
GROUP CDATA #IMPLIED PERMISSIONS CDATA #IMPLIED>
```

图 2 powermart.dtd 定义

XML 具有许多重要的特性, 首先是可扩展性。正如 C++ 让使用者声明自己的类, XML 让使用者创建和使用他们自己的标记而不是 HTML 的有限词汇表。可扩展性是至关重要的, 企业可以用 XML 为电子商务和供应链集成等应用定义自己的标记语言, 甚至特定的行业一起来定义该领域的特殊的标记语言, 作为该领域信息共享与数据交换的基础。

其次是灵活性, XML 提供了一种结构化的数据表示方式, 使得用户界面分离于结构化数据。在 XML 中, 可以使用样式表, 如 XSL (Extensible Stylesheet Language, 可扩展样式语言) 和 CSS2 (Cas-

cading Style Sheets Level 2, 层叠样式表第 2 进阶), 将数据呈现到浏览器中。另外, XML 文档之间的超链接功能由独立 XLink (Extensible Linking Language, 可扩展链接语言) 来支持, 所有这些方面都可以互相独立地改进并发展。

再者是自描述性, XML 文档通常包含一个文档类型声明, 因而 XML 文档是自描述的。XML 文档中的数据可以被任何能够对 XML 数据进行解析的应用所提取、分析、处理, 并以所需的格式显示, XML 表示数据的方式真正做到了独立于应用系统。

除此之外, XML 采用 Unicode 字符编码系统, 从而支持世界上几乎所有的主要语言, 并且不同语言的文本可以在同一文档中混合使用。

2 ETL 工具 Informatica PowerMart

ETL (Extract Transform Loading) 是数据从业务系统抽取、转换、转载到数据仓库的过程^[4], 是构建数据仓库最困难的部分。在具体的设计过程中, 大部分要解决的难题都集中在这一层。因此, ETL 工具的选择就显得尤为关键, 其运行性能, 支持并行处理的能力以及系统的稳定性都非常重要。

无疑在众多的 ETL 工具中, Informatica 公司的 ETL 工具 PowerMart 无论在执行能力还是战略远见的完整性上都是名列前茅的, 但它并不符合 CWM 规范^[5], 如果抽取工具换了其他的厂商, 要将 ETL 规则从 PowerMart 中导出, 再导入到新工具, 实现元数据的平稳交换并不是一件容易的事情。幸运的是, Informatica 公司开发的 PowerMart 的 DTD 正是 PowerMart 元模型的一种表示形式, 整个 ETL 规则 (XML 文档) 的语法正是由此定义的, 通过它可以剖析 PowerMart 元模型的组织结构。

2.1 Powermart.dtd 的组织结构

PowerMart.dtd 的组织结构定义自以下文档:

```
<! -- Informatica Repository DTD Grammar - Version 6.1 --
->
<! ELEMENT POWERMART(REPOSITORY * )>
<! ATTLIST POWERMART
```

```
CREATE_DATE CDATA #IMPLIED
REPOSITORY_VERSION CDATA #IMPLIED>
```

如上定义,整个 POWERMART 元素由任意多个 (0...*) REPOSITORY 子元素组成,含有两个属性 CREATE_DATE 与 REPOSITORY_VERSION。它们均为字符数据属性类型,都是可选的,而且没有缺省值。

```
<! ELEMENT REPOSITORY ( FOLDER * ) >
<ATTLIST REPOSITORY
NAME CDATA #REQUIRED
DATATYPE CDATA #IMPLIED
CODEPAGE CDATA #IMPLIED
VERSION CDATA #IMPLIED
PARENTREPOSITORY CDATA #IMPLIED>
```

一个 REPOSITORY 元素由任意多个 (0...*) FOLDER 子元素组成,含有属性 NAME、DATATYPE、CODEPAGE、VERSION、PARENTREPOSITORY。它们均为字符数据属性类型,都是可选的,而且没有缺省值。

```
<! ELEMENT FOLDER ( FOLDERVERSION *, CONFIG *,
SCHEDULER *, TASK *, SEESION *, WORKLET *, WORK-
FLOW * ) >
<! ATTLIST FOLDER
NAME CDATA #REQUIRED
DESCRIPTION CDATA #IMPLIED
SHARED ( SHARED | NOTSHARED ) #IMPLIED
OWNER CDATA $ IMPLIED
GROUP CDATA #IMPLIED
PERMISSIONS CDATA #IMPLIED>
```

一个 FOLDER 元素由任意多个 (0...*) FOLDERVERSION、CONFIG、SCHEDULER、TASK、SEESION、WORKLET、WORKFLOW 子元素组成,含有属性 NAME、DESCRIPTION、SHARED (属性值为 SHARED 或 NOTSHARED)、OWNER、GROUP、PERMISSIONS。它们均为字符数据属性类型,都是可选的,而且没有缺省值。其中,除了 FOLDERVERSION 之外的诸元素均用于描述整个数据仓库信息流的运行、实现及调度,对应自 CWM 管理层中的数据仓库处理包和操作包,与 ETL 规则本身关系不大。

2.2 Informatica PowerMart 的体系机构

PowerMart 是基于引擎和元数据驱动的体系架构,PowerMart 体系结构中有两个 Server:

Informatica Repository Server 和 Informatica Server。Informatica Repository Server 是资料库 Server,可用来管理 ETL 过程中产生的元数据,Informatica Server 则是 ETL 引擎^[6]。

客户端 PowerMart Designer 是设计开发环境,定义源及目标数据结构,设计数据转换规则,生成 ETL 映射;Workflow Manager 合理地实现复杂的 ETL 工作流,实现基于时间、事件的作业调度;Workflow Monitor 监控 Workflow 和 Session 运行情况,生成日志和报告;Repository Server Administrator Console 建立和维护资料库;Repository Manager 则定义了对资料库和安全性的管理。

在资料库已创建的前提下,通过客户端 PowerMart Designer 和 Workflow Manager 开发实现将数据从业务系统 ETL 到数据仓库的过程。

3 Designer 和 Workflow 的 XML 描述

Designer 通过 Mapping 生成 ETL 映射,一个完整的 Mapping 应包含了 Targets、Sources 以及 Transformation,在 Mapping 中 Mapplets 的应用也是非常的广泛。在 Mapping 的设计中通过各种 Transformation 规则(Filter、Join、Aggregator、Router 等)的应用,完成数据从 Targets 到 Sources 的 ETL。Mapping 的 XML 定义如下:

3.1 声明 Powermart.dtd

```
<? xml version = "1.0" encoding = "gb2312"? >
<! DOCTYPE POWERMART SYSTEM "powmart.dtd">
<POWERMART CREATION_DATE = "09/10/2009 11:05:11"
REPOSITORY_VERSION = "133.41">
```

3.2 定义 Repository, DatabaseType 为 DB2

```
<REPOSITORY NAME = "REP_NAME" VERSION = "133"
CODEPAGE = "MS936" DATATYPE = "DB2">
</REPOSITORY>
```

3.3 在 Repository 中又定义 Folder

```
<FOLDER NAME = "FLD_NAME" GROUP = "Administrators"
```

```
OWNER = " Administrator" SHARED = " NOTSHARED" DE-
SCRIPTION = "" PERMISSIONS = " rwxrw - r - - " >
</FOLDER>
```

3.4 定义 Source 和 Target

```
<SOURCE NAME = " SRC_NAME" DBDNAME = " DB_NAME"
OWNERNAME = " DB_OWNER" DESCRIPTION = "" BUSINESS-
NAME = "" DATABASETTYPE = " DB2" OBJECTVERSION = " 1" >
</SOURCE>
<TARGET NAME = " TRG_NAME" CONSTRAINT = "" DE-
SCRIPTION = "" BUSINESSNAME = "" DATABASETTYPE = "
DB2" TABLEOPTIONS = "" OBJECTVERSION = " 1" >
</TARGET>
```

3.5 最后定义了数据转换规则,实现 ETL 映射

```
<MAPPING NAME = " M_P_NAME" ISVALID = " YES" DE-
SCRIPTION = "" OBJECTVERSION = " 1" >
</MAPPING>
```

Mapping 只定义源及目标数据结构,设计数据转换规则,生成 ETL 映射,ETL 工作的执行和调度还需要在 Workflow 中完成。Workflow 中定义了 Session 以及 ETL 作业调度规则。每个 Session 都需关联到一个 Mapping。Session 可设置为共享与非共享,不同类型的 Session 在 XML 中定义的位置也不同。共享的 Session (REUSABLE = " YES") 在 { FOLDER } 中定义,非共享的 Session (REUSABLE = " NO") 则在对应的 { WORKFLOW } 中定义:

```
<FOLDER>
<SESSION NAME = " S_M_RNAME" ISVALID = " YES" REUS-
ABLE = " YES" SORTORDER = " Binary" DESCRIPTION = ""
MAPPINGNAME = " M _ RNAME" MAPPINGVERSION = "
1. 0. 0" ></SESSION>
<WORKFLOW>
<SESSION NAME = " S_M_P_NAME" ISVALID = " YES" REUS-
ABLE = " NO" SORTORDER = " Binary" DESCRIPTION = ""
MAPPINGNAME = " M _ P _ NAME" MAPPINGVERSION = "
1. 0. 0" ></SESSION>
</ WORKFLOW >
</FOLDER>
```

Workflow 则定义了 ETL 的调度规则:

```
<WORKFLOW NAME = " WF_P_NAME" ISVALID = " YES"
ISENABLED = " YES" SERVERNAME = " SER _ NAME" DE-
SCRIPTION
= "" SCHEDULERNAME = " Scheduler" SUSPEND_ON_ERROR
```

```
= " NO" REUSABLE_SCHEDULER = " NO" >
</WORKFLOW>
```

4 应用分析

广东发展银行 EDS 平台是满足分行、部门或其他系统的数据抽取需求,同时具有数据交换、整合服务功能,为各后台信息管理系统提供统一、整合的数据的全行数据应用的总枢纽。EDS 平台的数据分发 workflow 共有三百多个,ETL 逻辑分为实时同步和批量抽取两类,工作流的调度规则均类似。在 Workflow Manager 中开发一个实时同步逻辑的工作流平均需要 30 分钟(不包括设计阶段,只计生产时间),掌握了 PowerMart. dtd,则可以开发出一个通用的 XML 模版,针对不同的工作流,通过分析 Workflow 的 XML 描述可以看出只需对 XML 模版进行少许的修改,重指定对应的 Mapping,生成新的 XML 文件,再将该 XML 文件导入 PowerMart Repository 中,即可生成新的工作流作业。这种基于 XML 的开发方式极大地节省了开发时间,同时也可避免在 Workflow Manager 上执行繁琐的开发操作,结合 UltraEdit 来修改 XML,可更进一步提高开发效率。图 3 为基于 XML 技术开发的表 MCCADPF 的分发工作流。

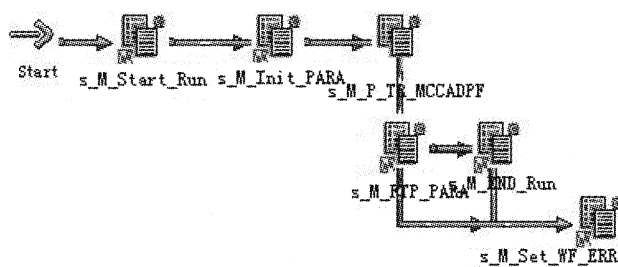


图 3 MCCADPF 表的分发工作流

5 小结

通过对 PowerMart. dtd 的分析和研究,提出了一种基于 XML 技术的开发方案,并且在广东发展银行 EDS 项目的开发中获得了成功,但这种基于 XML 的

(下转第 283 页)

A Method of Session Construction Based on Time-and-reference Heuristic

WANG Gai-xiang

(Shanxi Professional College of Finance, Taiyuan 030008, P. R. China)

[**Abstract**] Web usage mining is a popular research topic which combines various technologies and methods between data mining and WWW. It is that the data pretreatment is the foundation, the mining access pattern is the core, and the pattern analysis and demonstration is a goal in Web usage mining. In the procedure of data pretreatment, time-and-reference heuristic method is used to construct session, the method which is time-based heuristic and reference-based heuristic. The method not only uses the time characteristic of session between user session and Web sites, but also considers the users' navigating characteristic.

[**Key words**] Web usage mining time-and-reference time-based reference-based

(上接第 272 页)

开发方案有一定的局限性,需要结合项目的特点来开发出 XML 模版,该方法适合互用性较高的批量工作流的开发。

参 考 文 献

- 1 瞿裕忠,张剑锋,陈 峥,等. XML 语言及相关技术综述. 计算机工程,2000;12,(12):4—6
- 2 刘 绚,薛 贺. 基于 XML 的数据集成应用架构中的模式管理模型. 科学技术与工程,2007;7(3):325—328

- 3 王 悦. 基于 CWM 规范的异构平台抽取工具的元数据交换. 北京邮电大学学位论文,2005:42—54
- 4 William H. Frmon. 数据仓库. 王志海,译. 北京:北京机械工业出版社,2006
- 5 雷启明. 基于 CWM 的商场数据仓库 ETL 系统构架研究. 商业研究,2008;(551):39—40
- 6 杨晓东,姜 炜. Informatica PowerCenter V7. 1. 2 功能特性介绍. 北京:电子工业出版社,2005

XML Technology in Project Developing Based on PowerMart

WU Jun, HAN Bin, LIU Gang¹

(School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, P. R. China;

Guangzhou Highjet Computer Technology Ltd¹, Guangzhou 510665, P. R. China)

[**Abstract**] PowerMart is a powerful ETL tool, but during application developing process, it was founded that the interface of PowerMart was somewhat complicated, and the process was time-consuming, with full analysis of powermart. dtd, a new developing technique based on xml was proposed on this article, and for similar to ETL mapping or Workflow design a unified XML templates, by modifying the XML templates to generate a new ETL mapping or Workflow. The development of Guangdong Development Bank project EDS proves that this method can significantly improve develop efficiency.

[**Key words**] PowerMart. dtd XML technology extract transform loading