基于 XML 的数据转换系统 SuperETL*

柴 胜[1,2]，周云轩[2,3]，黄永平[1]，王洪媛[1]，王云霄[1]

(1. 吉林大学 计算机科学与技术学院，吉林 长春 130012；2. 吉林大学 地球探测科学与技术学院，吉林 长春 130026；3. 华东师范大学 河口海岸国家重点实验室，上海 200062)

摘 要：针对政府机构和企事业单位对数据资源整合的需求，提出一个数据转换系统 SuperETL，主要介绍其设计目标、体系结构，并给出了系统中任务的 XML 定义标准。测试结果表明，SuperETL 能够高效、智能地完成数据抽取(Extract)、清洗(Cleaning)、转换(Transformation)、装载(Loading)及 ETL 任务。

关键词：抽取；清洗；转换；装载

中图法分类号：TP311.52 **文献标识码**：A **文章编号**：1001-3695(2006)01-0016-03

Data Exchange System SuperETL Based on XML

CHAI Sheng[1,2], ZHOU Yun-xuan[2,3], HUANG Yong-ping[1], WANG Hong-yuan[1], WANG Yun-xiao[1]

(1. College of Computer Science & Telenology, Jilin University, Changchun Jilin 130012, China; 2. College of Faculty of GeoExploration Science & Technology, Jilin University, Changchun Jilin 130026, China; 3. State Key Laboratory of Estuarine & Coastal Research, East China Nomal University, Shanghai 200062, China)

Abstract: Based on the needs of data resources exchange and integration by government organizations and industries, this paper initiates a data exchange system SuperETL. Its design aims and system architecture are introduced. The XML criterion of the task in the system is defined. As a result, the SuperETL can implement intelligently the task of extract, transformation and loading(ETL).

Key words: Extract; Cleaning; Transformation; Loading

随着信息化进程的推进，政府机构和企事业单位对数据资源整合的需求越来越明显，越来越多的单位将数据集中纳入到下一步规划的重点。但面对分散在不同地区、种类繁多的异构数据库进行数据集中并非易事，首先要解决冗余、歧义等脏数据的清洗问题，仅靠手工进行不但费时费力，质量也很难保证；另外数据的定期更新也存在困难。如何实现业务系统数据集中是摆在政府机构、企事业单位面前进一步提升信息化程度的最大难题。

SuperETL 数据转换系统为数据大集中提供了令人满意的解决方案，它可以智能地批量完成数据抽取(Extract)、清洗(Cleaning)、转换(Transformation)、装载(Loading)[1]，即 ETL 任务，不但满足了用户对种类繁多的异构数据库进行整合的需求，同时可以通过增量方式进行数据的后期更新，一体化地解决了数据集中过程中遇到的种种困难。

1 SuperETL

1.1 设计目标

(1) 数据资源配置工具。对数据库操作的时候，系统通过这个工具进行数据库连接，可以直接配置 ODBC 支持的数据库(包括 Oracle, SQL Server 等)的连接。

(2) 任务配置工具。将系统已有的功能节点和拓展节点进行组合，形成执行流程，产生用户需要一次性完成的任务，同时提供新建、修改、删除任务的功能。

(3) 数据转换引擎。解释执行在任务配置中定义的各种任务。

(4) 调度配置。自动配置各个任务执行的时间。

(5) 调度控制。启动和停止调度服务。

(6) 任务列表。将所有任务展现给用户，用户可以手工进行执行操作。

(7) 任务监控。查看任务执行结果的成败。

(8) 日志查看。系统运行的日志信息，按照每次执行的任务进行详细处理结果列表。

1.2 体系结构

如图 1 所示，首先将待操作的数据源(即源数据源)通过数据源配置工具配置数据库的相关参数进行数据库连接。数据源配置好后为数据转换任务配置及执行任务调用做准备。任务配置实质是元数据(描述数据的数据)的生成和管理，可以通过可视化任务配置工具配置任务实现所需要用到的功能节点以及设定任务中各个功能节点的执行流程。配置好的任务将在任务列表中列出，可以手动执行，也可以由调度配置统一执行，通过调度控制来控制服务的启动和停止。任务的执行是由数据转换引擎解释处理的，同时记录明细日志，通过任务监控反映出任务整体执行情况，并通过日志查看具体任务节点的执行情况。

收稿日期：2004-09-29；修返日期：2005-03-23

基金项目：国家“863”计划资助项目(2003AA118020)；教育部高等学校优秀青年教师教学科研奖励计划资助项目

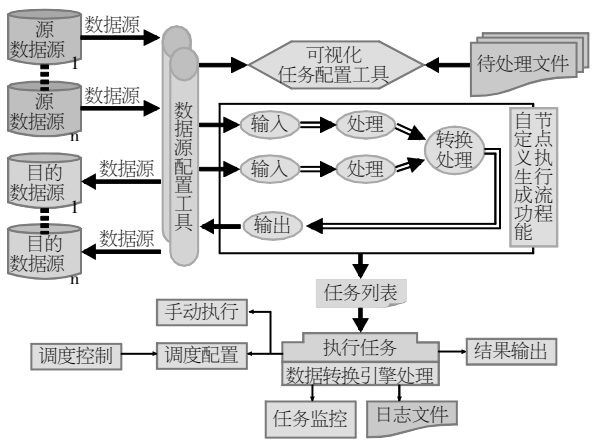

图 1　系统结构

2　核心原理与实现

SuperETL 数据转换系统的核心是任务的配置以及任务的执行, 前者是相关元数据的生成和管理, 后者是由数据转换引擎解释执行任务[2]。

2.1　任务的配置[5]

下面是几个相关概念:

(1) 数据变换是数据从源数据源到目标数据源的变换过程, 整个过程由任务描述。

(2) 任务由节点和流程构成, 由 XML 文件描述。

(3) 节点用来完成一些变换操作, 如数据的清洗、转换。

(4) 流程用来连接节点并传递数据。

每个节点有 0 ～*n* 个输入端口(InputPort) , 0 ～*n* 个输出端口(OutputPort) 。每条流程连接两个节点的输入与输出端口, 负责将数据从输入端口传输到输出端口。

一个任务至少有一个入口节点。它只有输出端口, 没有输入端口, 这个节点成为任务的入口点, 表示外部数据从此进入处理过程。相反地, 终端节点是没有输出端口的节点, 表示此节点保存数据。其余节点既有输入端口又有输出端口。

概念示意图如图 2 所示。

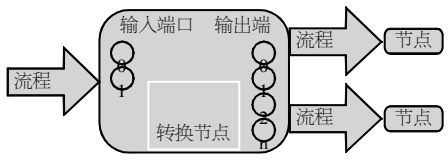

图 2　节点与流程示意

整个系统采用 Java 语言实现, 其中每个节点相当于 Java 语言中的一个类。系统目前已经实现了很多基础节点, 如数据库连接节点、增量抽取、增量载入、数据更新等。用户使用该系统时可以根据具体情况选择直接使用已有节点或者扩展相关节点完成自己特殊的需要。

对于系统中任务以及数据记录的定义, 我们提出了一种有效的标准, 该标准采用 XML 描述[3]。下面给出一个实际的例子:

(1) SuperETL 用 XML 定义数据记录格式规格, XML 文件扩展名为. fmt。

XML 的 DTD 如下:

```
<!ELEMENT Record( FIELD) >
<!ATTLIST Record
  name 唯一标志
  type ( fixed | delimited ) 定义类型 >
<!ELEMENT Field >
<!ATTLIST Field
  name 值为字段名唯一
  type 值为字段类型
  delimiter 值为定界符号, 可选, 默认“, ”
  size 值为长度, 可选, 默认“0 ”
  format 值为变换数据时格式限制, 可选
  nullable 值为是否允许包括 Null 值, 可选, 默认“True ”
  default 值为数据字段默认值 >
```

在下面例子中, 描述了包含三个字段的记录元数据:

```
< ?xml version = " 1. 0"  encoding = " UTF-8" ?  >
< Record name = " TestInput"  type = " delimited" >
< Field name = " Name"  type = " string"  delimiter = " ; " / >
< Field name = " Age"  type = " numeric"  delimiter = " |" / >
< Field name = " City"  type = " string"  delimiter = " \n" / >
< /Record >
```

记录有三个字段: Name (String 类型) ; Age (Numeric 类型) ; City (String 类型) 。

(2) SuperETL 中的任务仍然用 XML 格式定义元数据、节点和流程, XML 文件扩展名为. grf。

XML 的 DTD 如下:

```
<!ELEMENT Graph( Global, Phase) > 根元素( 包含 Global 和 Phase 子元素)
<!ATTLIST Graph name 名称, 唯一标志 >
<!ELEMENT Global( Metadata, DBConnection) > Global 的子元素在任务中为全局数据
<!ELEMENT Metadata > 格式描述元素
<!ATTLIST Metadata
  id 唯一标志
  fileURL 格式文件名 >
<!ELEMENT DBConnection > 数据连接元素
<!ATTLIST DBConnection
  id 唯一标志
  dbDriver JDBC 驱动程序
  dbURL 数据库连接 URL
  user 用户名
  password 密码 >
<!ELEMENT Phase( Node, Edge) > 分阶段描述节点与流程( 包含 Node 和 Edge 元素)
<!ATTLIST Phase
  number 顺序号, 唯一标志 >
<!ELEMENT Node >
<!ATTLIST Node
  type 节点类的类型
  id 唯一标志 >
<!ELEMENT Edge >
<!ATTLIST Edge
  id 唯一标志
  metadata 指向格式文件
  fromNode 起始节点
  toNode 终止节点 >
```

在 Edge 的 fromNode 和 toNode 的属性值应遵循如下样式:

< Node ID > : < Port Number > 即节点的 ID: 端口号

在下面例子中, 描述了数据对拷的任务

```
< Graph name = " MyTransformation" >
< Global >
< Metadata id = " DataTypeA"  fileURL = " $ HOME/ myMetadata/ dataTypeA. fmt" / >
< Metadata id = " DataTypeB"  fileURL = " $ HOME/ myMetadata/ dataTypeB. fmt" / >
< /Global >
< Phase number = " 0" >
< Node id = " INPUT"  type = " DELIMITED_DATA_READER"  fileURL = " c: \projects \jetel \pins. ftdglacc. dat"  / >
< Node id = " COPY"  type = " SIMPLE_COPY" / >
```

```
< Node id = " OUTPUT" type = " DELIMITED_DATA_WRITER" append = " false" fileURL = " c: \projects \jetel \pins. ftdglacc. dat. out" / >
< Edge id = " INEDGE" fromNode = " INPUT: 0" toNode = " COPY: 0" metadata = " InMetadata" / >
< Edge id = " OUTEDGE" fromNode = " COPY: 0" toNode = " OUTPUT: 0" metadata = " InMetadata" / >
< /Phase >
< /Graph >
```

MyTransformation 任务定义了 DataTypeA, DataTypeB 格式数据和 INPUT, COPY, OUTPUT 节点及 INPUT 到 COPY 再到 OUTPUT 的流程。通过格式文件 dataTypeA. fmt 从类型为 DELIMITED_DATA_READER 的节点读出数据,经过类型为 SIMPLE_COPY 的节点处理,以 dataTypeB. fmt 格式输出给类型为 DELIMITED_DATA_WRITER 的节点的一个数据对拷的转换过程。(DELIMITED_DATA_READER, SIMPLE_COPY, DELIMITED_DATA_WRITER 都是系统实现的节点类型,这里直接使用即可)。

2.2 任务的执行

第 2.1 节是任务的配置过程,即元数据的生成和管理。它可以手工完成,也可以通过 SuperETL 提供的可视化任务配置工具完成。用户定义完 SuperETL 系统的任务后,必须启动数据转换引擎解释执行该任务。数据转换引擎使用 Java 语言编写,引擎首先读取第 2.1 节定义的 XML 文件,使用 Java 语言中的 XML API 对它进行解释,然后以线程[4] 的方式执行每一个节点,引擎要监控所有节点的运行过程。

3 结束语

结合信息化过程中面临的 ETL 任务,提出并实现了基于 XML 的数据转换平台 SuperETL。主要做了以下工作(限于篇幅,SuperETL 系统其他部分没有详细介绍):①给出了 SuperETL系统的设计目标和体系结构;②给出了实现 SuperETL 系统的核心原理和实现,即任务的定义规范。进一步工作包括:①根据需要扩展更多的节点;②提高系统的稳定性和运行效率。

参考文献:

[1] iawei Han. Data Mining: Concepts and Techniques[M]. Beijing: Machine Industry Press, 2001.

[2] Tony Bain. Professional SQL Server 2000 Data Warehousing with Analysis Services[M]. Wrox, 2001.

[3] Brett McLaugblin. Java & XML (2nd Edition)[M]. O'Reilly & Associates, 2002.

[4] Scott Oaks. Java Threads[M]. America: O'Reilly, 2001.

[5] Lou Agosta. The Essential Guide to Data Warehousing[M]. Prentice Hall/Pearson, 2001.

作者简介:

柴胜(1976-),男,黑龙江人,讲师,博士生,主要研究方向为数据仓库、软件工程;周云轩(1962-),男,江苏睢宁人,博士生导师,主要研究方向应用地球物理和地理信息系统;黄永平(1964-),男,吉林人,副教授,主要研究方向为网络;王洪媛(1974-),女,吉林人,讲师,博士生,主要研究方向为软件工程;王云霄(1976-),女,河北人,讲师,博士生,主要研究网络安全、图像处理。

(上接第 11 页) 不应忽视人的作用,因为应急系统永远不可能代替人在处理应急事件时所体现出的经验和智慧。

4 结束语

人为(如战争、恐怖事件)和自然(如地震、海啸)灾害会造成巨大的破坏,涉及社会生活和经济建设、国家安全的各个方面,应急系统对于减灾、救灾会发挥重要作用,所以应该十分重视。目前应急系统普遍性原理和一般性理论还远未成型,因而在应急系统设计开发中人的作用不容忽视,人机结合应急系统研究值得提倡。另外应急系统研究多局限于应急管理和决策系统,而对应急设计系统较少见,尤其是缺乏应急设计理论指导[16]。特定的应急系统研究与开发应该针对该灾难的特点,吸取各行业应急系统的开发设计理论、技术方法、实践经验成果,并在理论与实践相结合中不断地完善和发展。

参考文献:

[1] 翟晓敏,盛昭瀚,何建敏.应急研究综述与展望[J].系统工程理论与实践,1998,(7):17-24.

[2] 陈永安.当前政府建立应对突发事件应急管理系统的思考[J].云南行政学院学报,2003,(4):20-23.

[3] 袁辉.重大突发事件及其应急决策研究[J].安全,1996,(2):1-4.

[4] T Elbir. A GIS-based Decision Support System for Estimation, Visualization and Analysis of Air Pollution for Large Turkish Cities[J]. Atmospheric Environment, 2004, 38(27): 4509-4517.

[5] H X Lan, C H Zhou, L J Wang, *et al*. Landslide Hazard Spatial Analysis and Prediction Using GIS in the Xiaojiang Watershed, Yunnan, China[J]. Engineering Geology, 2004, 76(1-2): 109-128.

[6] A Zerger, D I Smith. Impediments to Using GIS for Real-time Disaster Decision Support[J]. Computers, Environment and Urban Systems, 2003, 27(2): 123-141.

[7] M Hadjimichael, A P Kuciauskas, L R Brody, *et al*. MEDEX: A Fuzzy System for Forecasting Mediterranean Gale Force Winds[J]. Fuzzy System, 1996, (1): 529-534.

[8] P Avesani, A Perini, F Ricci. Interactive Case-based Planning for Forest Fire Management[J]. Applied Intelligence, 2000, 13(1): 41-57.

[9] J Z Hernández, J M Serrano. Knowledge-based Models for Emergency Management Systems[J]. Expert Systems with Applications, 2001, 20(2): 173-186.

[10] S P Kourniotis, C T Kiranoudis, N C Markatos. A Systemic Approach to Effective Chemical Emergency Management[J]. Safety Science, 2001, 38(1): 49-61.

[11] J L Wybo, K M Kowalski. Command Centers and Emergency Management Support[J]. Safety Science, 1998, 30(1-2): 131-138.

[12] Kevin F R Liu. Agent-based Resource Discovery Architecture for Environmental Emergency Management[J]. Expert Systems with Applications, 2004, 27(1): 77-95.

[13] R Granlund. Web-based Micro-World Simulation for Emergency Management Training[J]. Future Generation Computer Systems, 2001, 17(5): 561-572.

[14] A M Schaafstal, J H Johnston, R L Oser. Training Teams for Emergency Management[J]. Computers in Human Behavior, 2001, 17(5-6): 615-626.

[15] 余雁.国外化学危险品应急反应系统简述[J].安全,2002,(2):42-44.

[16] 张宝,滕弘飞.机械应急系统研究[J].中国机械工程,2003,14(10):888-891.

作者简介:

郑晓军(1982-),男,山西人,博士研究生,研究方向为计算机智能、应急系统;王奕首(1978-),男,博士研究生,研究方向为布局优化、应急设计、计算智能;滕弘飞(1936-),男,教授,博士生导师,研究方向为智能 CAD、计算智能、人机结合、布局。