一种轻量级电网实时数据ETL系统的设计与实现

段 成[1,2]，王增平[1]，吴克河[2,3]

（1.华北电力大学电气与电子工程学院，北京 102206；2.北京市电力信息技术工程研究中心，北京 102206；
3.华北电力大学控制与计算机工程学院，北京 102206）

摘要：以某省电网实时数据监测系统的数据中心建设项目为背景，提出了一种轻量级的，适合于电网实时数据迁移与同步更新的ETL系统解决方案。系统支持多种关系型数据库和数据文件，提供全量、增量、主细表等多种自定义模式数据迁移方案。系统利用JDBC数据库访问技术、JSR-166的Util.Concurrent并发线程包以及SWT技术，结合元数据的管理，解决了异构数据的快速抽取、清洗转换与加载、任务的调度和跨平台运行的问题，并在实际应用中表现出了实用性和稳定性。

关键词：ETL系统；异构数据；增量；任务调度；跨平台

Design and implementation of a lightweight ETL system for power real-time data

DUAN Cheng[1,2]，WANG Zeng-ping[1]，WU Ke-he[2,3]

（1. College of Electrical & Electronic Engineering，North China Electric Power University，Beijing 102206，China；
2. Beijing Electric Power Information Technology Research Center，Beijing 102206，China；
3. College of Control & Computer Engineering，North China Electric Power University，Beijing 102206，China）

Abstract：Using the data center construction project for a power real-time monitoring system as the background，this paper proposes a novel lightweight ETL solution for the migration and synchronization of power real-time data. The solution not only supports a variety of relational database and text-based data sources，but also provides total，incremental，main-sub table and several other custom mode data migration schemes. By combining JDBC database access technology，Util.Concurrent package of JSR-166 and SWT technology with the management of metadata，our system solves the problems of heterogeneous data extraction，cleaning，transformation and loading，task scheduling and cross-platform operation. Practical applications demonstrate the practicality and stability of this system.

Key words：ETL system；heterogeneous data；incremental；task scheduling；cross-platform

中图分类号：TM76 文献标识码：A 文章编号：1674-3415(2010)18-0174-04

0 引言

随着信息技术的蓬勃发展，各供电公司为了保障电网安全、可靠、稳定、经济的运行，针对不同应用需求建立了各类业务系统，相关的数据信息是多源异构的，即数据的存储介质、存储形式、存储位置以及所处的网络条件都不尽相同。当前电网企业管理模式由分散化向集中化、精益化转变，对电网系统做全面深层次的分析就需要整合各类电网数据，建立数据中心以便为用户提供全局统一的数据视图。ETL是构建数据中心的重要环节[1]，在跨越多个平台的复杂网络条件下，依赖供应商的ETL产品往往难以满足电网实时数据迁移与同步的需要，这就需要实现一个独立的ETL系统。

在为某省电网公司建立电网实时数据监测系统数据中心的过程中，需要从盐密监测系统（Oracle），微气象系统（Access），蓄电池监测系统（Sqlserver），杆塔线路监测系统（Excel）等众多的数据源中获取实时数据。以往的做法是通过编写一系列专门的接口程序来实现的，为了降低此类项目的实施难度和开发成本，我们开发了一个独立的ETL系统来满足电网实时数据的迁移与同步更新的需要。

1 ETL及其实现相关技术背景

1.1 ETL系统介绍

ETL（Extraction Transformation Loading）是抽取-转换-加载的英文缩写，具体来说，是从各种类型的数据源中抽取数据，并对其进行清洗与转换，

最终将满足要求的数据内容装载入数据仓库的过程[2]。

抽取（Extraction）是从分布式的异构数据源中抽取业务数据的过程。数据源包括 Oracle，Sqlserver，Sybase，DB2，Access 等关系型数据库以及 Excel，Xml 等数据文件。

转换（Transform）是将抽取所得的数据经过一系列的变换加工转变为符合目标数据模式的过程，包括数据质量的检查，数据的清洗，数据的集成与合并等，同时，在此过程中使用的转换规则分别在元数据中得以描述和保存。ETL 数据转换过程如图 1 所示。

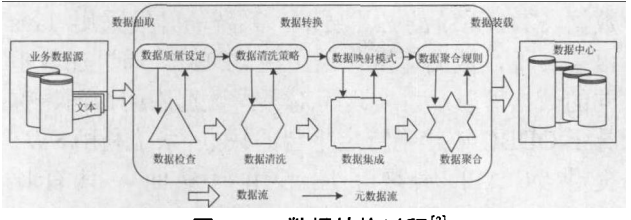

图 1 ETL 数据转换过程[3]

Fig.1 ETL process of data conversion

加载（Loading）是指将转换好的符合目标模式的数据载入数据仓库的过程[4]。

1.2 相关技术

数据库访问技术 JDBC：JDBC（Java Data Base Connectivity）由一组用 Java 语言编写的类和接口组成，是一种用于执行 SQL 语句的 Java API，可以为各种关系数据库提供统一的访问接口。

并发线程包：来自 JSR-166（Java Specification Requests）的 Util.Concurrent 并发线程包提供了一系列高性能并发程序开发的实用模型，已经成为 J2SE5.0 的一部分。

SWT 图形组件技术：SWT（Standard Widget Toolkit）是一个独立于平台的，可以脱离 Eclipse 框架单独使用的图形组件。SWT 通过 Java 本地接口（JNI）直接使用操作系统自身的各种图形控件，开发出来的应用程序具有本机应用程序的观感。

2 系统的总体架构与设计

2.1 系统的整体结构

电网实时数据通常是以关系型数据库表或者数据文件的形式存在，它们共同构成了企业的领域知识和行为规则库。本文所指的 ETL 系统是将这些数据进行处理并加载入数据仓库的过程。系统的构架图如图 2 所示。

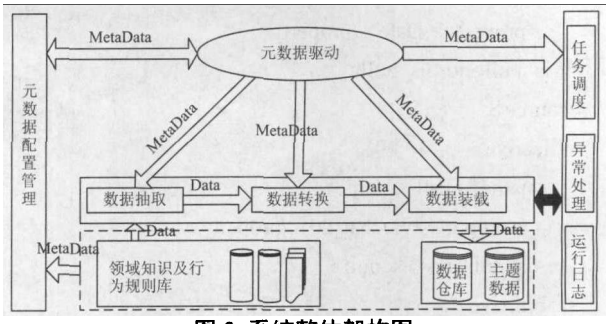

图 2 系统整体架构图

Fig.2 Overall system architecture diagram

系统按照功能划分为三个模块，分别是元数据管理模块、数据处理模块、任务调度及监控模块。

元数据管理模块：ETL 任务的生成及数据的处理过程都是基于元数据驱动的，元数据管理是 ETL 系统的重要组成部分，对元数据进行合理有效地管理，才能很好地实现数据的访问、清洗、转换与加载。元数据管理模块的功能是负责各种类型元数据的定义、存取及管理维护。

数据处理模块：数据处理模块是 ETL 系统的核心模块，系统通过 JDBC 连接各种数据源，从其中抽取数据，对数据进行清洗转换（数据的筛选，数据质量的校验，数据属性及域的转换，数据的合并等），最终将数据按照预先定义的目标模式装载入数据仓库之中。

任务调度及监控模块：鉴于数据的 ETL 是一个反复进行的过程，而且各类电网实时数据的数据密度也不尽相同，往往需要定义不同周期不同频度的作业任务，任务的调度保证了任务的自动、高效、准确的运行。此外针对 ETL 任务执行的过程中可能出现各种各样的异常，比如网络的中断，任务的超时等情况进行及时的处理，并支持断点续传的功能。系统可以对任务的全生命周期进行监控，并将其执行过程记录日志，运行日志的主要功能是记录任务执行过程中发生的操作以及错误、警告等信息，以便于在事后进行分析和跟踪处理。

2.2 元数据定义

元数据（Metadata）即关于数据的数据。系统中元数据定义主要包括数据库的驱动信息、用户名、密码、连接 URL、表、列的属性（类型，格式、约束条件等）、源数据到目标数据字段间的映射关系、数据的转换规则、任务的参数定义等等。以下给出了一个字段间映射关系 xml 片段：

```
<property-filedmap>
<field>
  <source>
    <name>YMD</name>
```

```
        <property> Date </property>
      <is_null>no</is_null>
    </source>
    <objective>
        <name>TIME</name>
          <property>varchar(20)</property>
        <is_null>no</is_null>
     </objective>
</field>
      ...
</property-filedmap>
```

上述 xml 片段所示的是将 YMD 字段映射到目标字段 Time 上，其中 YMD 字段属性是 Date 型，Time 字段属性域是 varchar(20)，两者都是非空的。数据处理程序是按映射规则将源数据转换为目标数据模式。

系统可以自动匹配字段间的映射关系，也能够人工地进行配置，可以为任务设定周期及参数。方案配置以 xml 文件的形式保存。图 3 是方案的配置界面。

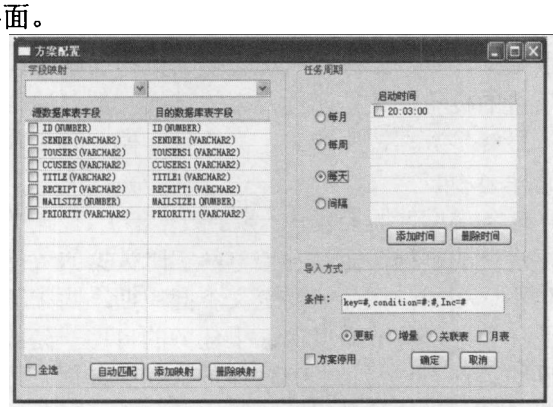

图 3 方案配置窗口

Fig.3 Program configuration window

2.3 并发任务的实现

ETL 系统是面向多任务的，多线程编程是其中的重点及难点，线程的互斥、同步事件的响应、跨线程的数据通信以及异步任务调度都是需要解决的问题。我们利用 Util.concurrent 并发线程包提供的锁、互斥、队列、线程池、轻量级任务等一系列的并发构件很好地解决了上述问题。

程序中用到了 Util.concurrent 包提供的 Executor、Callable 和 Future 等几个基本的接口。通过 Executor 接口提供的最基本的任务执行方法，实现了任务的 submit（）和 shutdown（）等操作；通过 Callable 接口获得任务抛出的异常和返回结果；使用 Future 接口获取任务的句柄，从而得到任务执行的中间结果，终止任务的执行等，可以很方便地对任务执行过程进行全生命周期的监控；此外还使用了 Util.concurrent 包提供的一些辅助工具类。

```
public static ScheduledThreadPoolExecutor scheduler;
scheduler.schedule(event,delay,TimeUnit.MILLISECONDS);
```

以上代码片段是将一个任务 event 提交给了线程池 ScheduledThreadPoolExecutor，由线程池来负责任务的调度，event 将在指定延时 delay 后得到执行。

2.4 跨平台运行

ETL 系统的开发和设计与计算机操作系统平台密不可分，电力系统出于安全和稳定性的考虑运行着各种平台，能够在多种平台下运行，是电网实时数据 ETL 系统的必然要求。系统的开发采用 Java 语言实现，数据库操作采用纯 JDBC，通过加载不同的数据库驱动，实现了对不同数据库的访问，弥补了 ODBC 依赖于特定平台的缺点。系统利用 SWT 提供的 API 实现了图形用户界面，具有比 AWT/Swing 应用程序更快的响应速度。

系统具有良好的可移植性，图 4 是部署在 Kylin Linux 上的盐密监测系统数据迁移方案实例，从图中可以看出系统具有 Linux 桌面应用程序的观感。

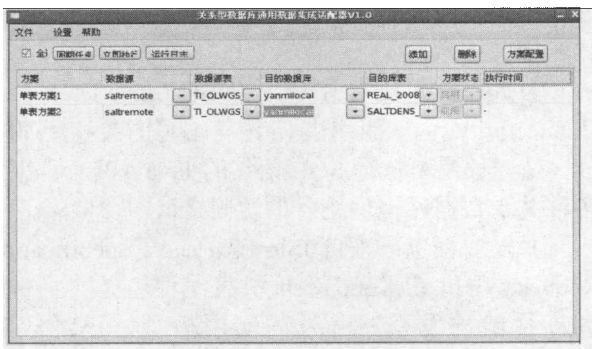

图 4 Linux 下运行效果

Fig.4 Running effects under Linux

3 数据迁移与同步模式

现行各类电网实时系统中的数据量动辄就成千上万条，其中既有动态的，也有静态的数据，而数据迁移与同步操作是一个定期的、频繁的工作，数据的更新频度可能是每月、每周、每天甚至是 5 s 这样一个很小的时间间隔，为了最大限度地利用系统资源，提高系统的性能，并满足实际应用的需要，系统实现了多种模式的数据迁移方案；同时为了方案定义的灵活性，本文提出了一种自定义的表达式，后台程序可以通过解析表达式中的关键字参数，生成代码参数或动态 SQL 语句[5]。具体实现方式如下：

① 全量模式：全量更新模式将源数据表中的数

据整体向目的表迁移，主要用于静态数据表的更新。根据实际应用需要实现了两种方式：一种是将目的表所有数据整体删除，然后再更新；另一种可以通过关键字段“逐条比对”的方式，实现数据整体同步更新，倘若关键字段值相同则更新记录，否则插入记录。后者效率较低，但能够保护已有的业务记录。全量模式表达式如下：

Total = :{key=#,Condition=#;#}

其中：“#”表示默认不设置情况； key 对应目的表的主键 ID 的生成规则，比如针对 Oracle 数据库可设定一个序列名 seq_salt ；Condition 为数据筛选条件，通过“；”分隔可设置多个条件。

② 增量模式：即保证上一次已经传输过的数据只要在此期间没有任何变动，下一次传输时将被忽略，当前的技术难以实现通用的增量数据抽取[6]。针对电网实时数据的特点，本文提出了增量标识字段的方式，实现了数据的增量迁移。标识字段可以是时间戳字段，ID 自增字段以及任何具有增量性质的字段。倘若数据源表不含具有增量性质的字段，则采用关键字段（主键）“逐条比对”的方式实现数据迁移，如果关键字段值在目的表中已存在则忽略此记录，否则插入新记录。增量模式表达式如下：

Incremental = :{ key=#,Condition=#;#,Inc=#}

式中：“Inc”用来标识增量字段，比如“Inc=YMD”，表示将“YMD”时间字段作为增量标识字段；任务执行完毕后，系统将记录“YMD”字段的当前值以供下次任务执行时生成 sql 语句。比如“select * from saltreal where YMD>2009-02-26 08:30:20”，其中“2009-02-26 08:30:20”就是系统记录的“YMD”字段的值。

③ 主细表模式：在数据库应用系统的开发过程中，往往用外码来保证参照实体间的数据一致性，这种有参照关系的表称为主细表。主细表模式可以保证数据的完整性，一致性，可用性，避免“脏”数据的产生，具有很高的实用意义。主细表模式表达式如下：

Main-sub = :{ Key=#,Inc=#,Condition=#;#

MT=#,SF=#:#;#:#,RF=#:#}

MT 对应的是主表名（静态信息表）；SF 对应源表字段与主表字段的映射关系；RF 对应主表字段与细表字段的映射关系；数据迁移执行过程中将根据 SF 的设置先查询主表，然后根据 RF 的设置将查询结果集映射到细表中。

为了提高系统的应用性能，实现了单表与多表两种数据迁移方案。单表方案是将对参数或更新频度有特殊要求的任务进行单独配置，任务对应一个独立的线程；多表方案是将更新频度相近且实时性要求不高的若干任务组成一个组方案，任务共享同一个线程顺序执行。具体实例在此不再详述。

4 应用实例

盐密实时监测系统中的实时监测表共有历史数据 63 000 余条，在方案配置时选用增量数据迁移模式，将 time 字段作为增量标识字段，任务执行周期设定为 1 h。该方案在小型机（IBM P5 570 H，8 G，4 CPU，Kylin Linux）上运行，首次执行需要迁移全部数据，共耗时 58 s，平均每秒处理 1 000 条以上记录；此后按增量方式运行，每次同步更新 6 条记录，系统已连续正常运行 6 个多月，实际应用表现出了很高的数据处理效率和稳定性。

5 结语

本文介绍了电网实时数据 ETL 系统实现的相关技术，阐述了系统的设计及实现方法。根据电网实时数据的特点，系统实现了多种模式的数据迁移方案，具备实用性与灵活性。ETL 是数据中心构建的重要环节，目前系统已经取得了实际的应用，并很好地满足了电网实时数据中心构建的需要，该系统对电力及其他行业数据中心ETL 的实施有一定的借鉴意义。

参考文献

[1] 白莉珍. ETL在青海省电力公司数据中心系统中的应用[J]. 青海电力，2008，27（2）：66-68.
BAI Li-zhen. Application of ETL in information center of Qinghai electric power company[J]. Qinghai Electric Power，2008，27（2）：66-68.

[2] Panos Vassiliadis，Alkis Simitsis，Spiros Skiadopoulos. Conceptual modeling for ETL processes[C]. //ACM SIGIR. Proc. of DOLAP'02. Virginia(USA), New York(USA)：2002：14 -21.

[3] Leon Gong，Mike Olivas，Christine Posluszny，et al. Deliver an effective and flexible data warehouse solution [EB/OL].[2005-08-04].http://www.ibm.com/developerworks/data/library/techarticle/dm-0508gong/index.html.

[4] 屈志毅，张延堂，王戈. 一种金融系统专用ETL工具的研究与实现[J]. 计算机工程，2008，34（20）：86-87，91.
QU Zhi-yi，ZHANG Yan-tang，WANG Ge. Study and implementation of special ETL tool for finance system[J]. Computer Engineering，2008，34（20）：86-87，91.

[5] 刘海英，冯文秀，杜晓通. 管理信息系统升级过程中数据迁移的研究及实现[J]，电力自动化设备，2005，25（5）：37-39.

（下转第 182 页　continued on page 182）

互感器模型。

4 结论

(1) 借助PSCAD自定义元件方法建立的具有小气隙环形铁芯的TPY级电流互感器模型调用简单，可通过灵活设置线圈匝数、铁芯截面积、气隙长度及二次负荷等各种相关参数调节互感器的性能，满足用户需要。

(2) 仿真计算实例表明该模型的合理性，该模型的建立丰富了PSCAD/EMTDC仿真程序的互感器元件模型，为基于PSCAD的继电保护仿真提供了前提。

在实际应用中，本文模型所需的基本参数要通过 CT 测试仪测量得到，这样才能保证仿真结果与实际相符。

参考文献

[1] 刘万顺，张忠理，杨奇逊，等. 电流互感器暂态过程的数字仿真[J]. 电路与系统学报，1996，1（1）：70-74.
LIU Wan-shun，ZHANG Zhong-li，YANG Qi-xun，et al. Digital simulation of current transformer transient response[J]. Journal of Circuits and Systems，1996，1（1）：70-74.

[2] 张军，朱声石. 电流互感器暂态仿真数学模型的改进[J]. 电力自动化设备，1997，17（3）：44-46.
ZHANG Jun，ZHU Sheng-shi. Improvement in math model for CT transient simulation[J]. Electric Power Automation Equipment，1997，17（3）：44-46.

[3] 余保东，张粒子，杨以涵，等. 电流互感器铁心的暂态磁化模型及误差计算[J]. 电工技术学报，1998，13（6）：25-30.
YU Bao-dong，ZHANG Li-zi，YANG Yi-han，et al. The transient magnetizing model of core and the error computation for CT[J]. Transactions of China Electrotechnical Society，1998，13（6）：25-30.

[4] 胡晓光，王哲，于文斌. 电流互感器暂态过程的仿真与分析[J]. 电力系统及其自动化学报，2001，13（4）：12-15.
HU Xiao-guang，WANG Zhe，YU Wen-bin. The transient simulation and analysis of current transformers[J]. Proceedings of the CSU-EPSA，2001，13（4）：12-15.

[5] James G Frame，Narendra Mohan，Liu Tsuhuei. Hysteresis modeling in an electromagnetic transients program[J]. IEEE Trans on PAS，1982，101（4）：3403-3411.

[6] Jiles D C，Atherton D L. Theory of ferromagnetic hysteresis[J]. Journal of Magnetism and Magnetic Materials，1986，61：48-60.

[7] Thoelke J B，Jiles D C，Devine M K. Numerical determination of hysteresis parameters for the modeling of magnetic properties using the theory of ferromagnetic hysteresis[J]. IEEE Trans on Magnetics，1992，28(1)：27-35.

[8] Annakkage U D，McLaren P G，et al. A current transformer model based on the Jiles-Atherton theory of ferromagnetic hysteresis[J]. IEEE Trans on Power Delivery，2000，15（1）：57-61.

[9] 阿法纳西耶夫 B B，等. 电流互感器[M]. 陆安业，等，译. 北京：机械工业出版社，1989.
Афанасьев В В，et al. Current transformer[M]. LU An-ye，et al，Trans. Beijing：China Machine Press，1989.

[10] Talukdar S N，Bailey J R. Hysteresis models for system studies[J]. IEEE Trans on PAS，1976，95(4)：1429-1434.

收稿日期：2009-09-23；　修回日期：2009-11-24

作者简介：

黄　莉（1978-），女，讲师，硕士，从事水电站计算机监控系统与综合自动化研究；E-mail：ee_hl@126.com

杨卫星（1979-），男，工程师，硕士研究生，从事继电保护研究；

张雪松（1979-），男，高级工程师，博士，从事继电保护研究。

（上接第 177 页　continued from page 177）

LIU Hai-ying，FENG Wen-xiu，DU Xiao-tong. Study and implementation of data transfer in management information system upgrade[J]. Electric Power Automation Equipment，2005，25（5）：37-39.

[6] Prabhu Ram，Lyman Do. Extracting delta for incremental data warehouse maintenance[C]. //IEEE TCDE. Proc. of ICDE'00，San Diego，IEEE Computer Society. CA(USA),Washington D C(USA)：2000：220-229.

收稿日期：2009-10-14；　修回日期：2010-04-22

作者简介：

段　成（1985-），男，博士生，主要从事电力信息技术、分布式智能软件技术方面的研究；E-mail：duancheng1985@126.com

王增平（1964-），男，教授，博士生导师，主要从事电力系统继电保护、自动化方面的研究；

吴克河（1962-），男，教授，主要从事电力智能软件技术、计算机网络安全方面的研究。