利用 DTS 组件实现数据仓库中 ETL 方案设计①

Using DTS Component to Realize ETL Scheme in Data Warehouse

邱云飞 （辽宁工程技术大学软件学院 辽宁葫芦岛 125100）
邵良杉 那宝贵 （辽宁工程技术大学系统工程研究所 辽宁阜新 123000）

摘要：针对企业集成异构信息系统中数据，构建数据仓库辅助决策的需求，本文根据现场的实际状况设计了一个三级的 ETL 方案，负责将原始的业务数据源中的数据按照不同的决策主题进行集成，进而加载到数据仓库中。在此基础上利用 DTS 组件实现了一个 B/S 模式下的异构数据源 ETL 方案，并且结合我们承担的张家口煤矿机械有限公司决策支持系统项目给出了该方案的总体结构和部分关键实现代码。

关键词：数据仓库 ETL DTS ASP

1 引言

当前，企事业单位大都建立了信息系统，在运行了较长时间后，积累了海量的历史数据，这些数据都是企业宝贵的信息资源，其中隐含着各种有价值的信息，它们不但描述了企业过去的运行状态，也从某种程度上预示着企业将来的发展方向。因此，有效集成这些信息系统中的数据，构建数据仓库系统，为经营者提供决策支持信息就成为信息工作者的重要工作。但是，由于企业的信息化是一个长期的过程，且大多是在缺乏总体规划的前提下由不同的公司，基于不同的操作系统、数据库管理系统开发和实施的，因此，由原有的异构环境下的 OLTP 环境向 OLAP[1] 环境转换就成为数据仓库构建过程中最重要的一步。本文以我们在张家口煤矿机械有限公司决策支持系统项目中集成异构数据源、构建数据仓库的实例给出了一个 B/S 模式下数据仓库 ETL 的整体解决方案。

2 ETL 子系统设计

2.1 ETL 原理

ETL 是数据抽取（Extract）、转换（Transform）、清洗（Cleaning）、加载（Loading）的过程。这是构建数据仓库的重要环节，用户从数据源抽取出所需的数据，经过数据转换、数据清洗，最终按照预先定义好的数据仓库模型，将数据加载到数据仓库中[2]。

具体来说，数据抽取是数据源接口从业务系统中抽取数据，为数据仓库输入数据。数据转换包含对来自多个生产系统的数据源的处理，保证数据按要求装入数据库。数据清洗，这是一个确保数据集中的所有数值是一致的和被正确记录的处理过程。数据加载部件负责将数据按照物理数据模型定义的表结构装入数据仓库。

2.2 ZMJ_DSS 项目的 ETL 子系统设计

本论文所提出的利用 DTS 组件实现数据仓库中 ETL 方案设计是作者在开发张家口煤矿机械有限公司决策支持系统（简称 ZMJ_DSS）项目下的 ETL 子系统中总结出来的，下面我们将介绍一下该子系统。

ETL 子系统负责完成数据从原始数据源向目标数据仓库转化的过程，是 ZMJ_DSS 的一个重要部分。该模块将原有业务系统数据库和外部数据源的数据按照企业信息模型整合到数据仓库系统中去，从而为决策支持系统前端的 OLAP 和 DW 做好充分的数据准备。该系统的总体架构如图 1 所示[3]。

Hterp 数据库、Camplan_db 数据库、cbdb 数据库、Hterp 人力资源数据库分别是张家口煤矿机械有限公司的业务系统（包括 HT－ERP 系统、俐玛财务管理系统、车间成本核算管理系统、人力资源管理系统）下的

① 本文得到教育部博士点基金（20041047006）、国家自然科学基金（70572070）资助

数据源，ZMJ_DSS 项目数据仓库中的数据都从这些数据源获取。ZMJ_ERP 是 ZMJ_DSS 项目数据仓库所用到的业务源数据表的镜像，该数据库里的数据主要由对 Hterp 数据库、Camplan_db 数据库、cbdb 数据库、Hterp 人力资源数据库中数据仓库用到的数据表统一格式复制、整理得到的。ZMJ_ODS 对应数据仓库中的 ODS（操作数据存储）概念部分，包含全局一致的、细节的、当前或接近当前的数据，可以进行全局联机操作型处理，同时它是一种面向主题的、集成的数据环境，数据量小，适用于辅助完成日常决策的数据分析处理，该数据库中的数据通过 ETL 工具从 ZMJ_ERP 中抽取相关业务数据获得。ZMJ_DSS 数据仓库是系统数据组织存储层的核心，通过 ETL 工具对 ZMJ_ODS 层数据的抽取/转换/融合而形成综合了统计元素库、包括了从细节级、轻度综合、中度综合直至高度综合各级粒度的数据层，是按照主题分析的需要建立的企业级全局数据存储。

3　基于 DTS 组件的数据仓库中 ETL 解决方案设计

3.1　方案设计框架

根据张家口煤矿机械有限公司信息化建设的特点，以及系统开发工具的特点，其原有的业务系统中的数据大多采用微软的 SQL SERVER、ACCESS 数据库管理系统存储，业务系统也都是由 VB 或 ASP 开发。为达到更好的系统兼容性和运行效果，本方案的数据 ETL 过程采用微软的 DTS[4] 通过标准接口 OLE DB 或 ODBC（针对不支持 OLE DB 的数据源，如：Sybase）定义 ETL 过程的数据源连接，通过 DTS 自带的抽取规则或使用 T－SQL 脚本语言定义数据抽取、清洗和转换方法，使用微软 SQL SERVER 的 DTS 工具设计并完成所有的数据仓库中的 ETL 操作。用 DTS 工具设计完 DTS 包后，可以对包进行一次性执行，也可以把包设置为自动调度，使包的执行过程无需人工干预。

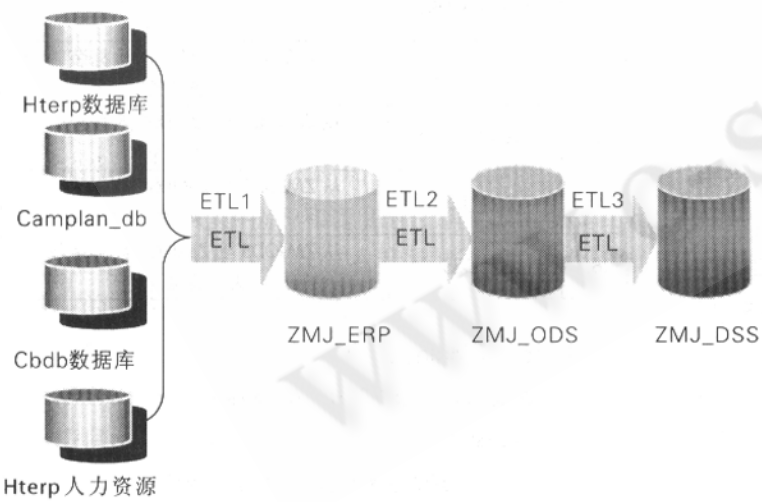

图 1　ZMJ_DSS 项目中 ETL 子系统总体架构

由图 1 可以看出，该 ETL 子系统一共有三个 ETL 过程，为方便起见，我们分别命名为 ETL1、ETL2 和 ETL3。三步 ETL 过程的详细说明见下表 1。

表 1　ETL 过程说明

ETL 过程	过程说明
ETL1	以实现数据共享、解决数据孤岛问题为目标，主要完成统一数据格式、清洗、转换和复制数据的操作，统一数据来源，实现异构数据库的类似 ERP 的集成，负责将原始的业务数据源中的数据加载到 ZMJ_ERP 中；
ETL2	以辅助企业日常决策为目标，按照不同的主题实现当前或接近当前数据的集成，构建适用于辅助完成日常决策的数据分析处理环境，负责将 ZMJ_ERP 中的当前或接近当前的数据加载到 ZMJ_ODS 中；
ETL3	以辅助企业战略决策为目标，完成数据的汇总和相应的转换工作，负责按照不同的主题将 ZMJ_ODS 中的数据增量加载到 ZMJ_DSS 中。

为了给系统管理员提供方便，将后台的 DTS 包的执行和调度通过 ASP 技术实现为 B/S 模式用户界面，这样系统管理员无须在服务器上对数据仓库的 ETL 进行管理和维护，管理员可以在其他任何一个地方（只要他能够访问到服务器）完成管理和维护操作。该方案的设计框架如下图 2 所示。

3.2　DTS 包设计

DTS 组件提供一组组件，可以从不同的数据源将数据抽取、转换和合并到一个或多个目标位置。借助 DTS，可以创建适合于各种特定需要的自定义数据转移解决方案。并且数据抽取工具需要一个能够统一访问多种数据源的数据访问接口，而基于 OLE DB 的 DTS 组件正具备这种能力，因此，使用 DTS 作为数据抽取工具开发的基础可以说满足了多方面的要求。

基于 DTS 组件，结合 ZMJ_DSS 项目中 ETL 子系统

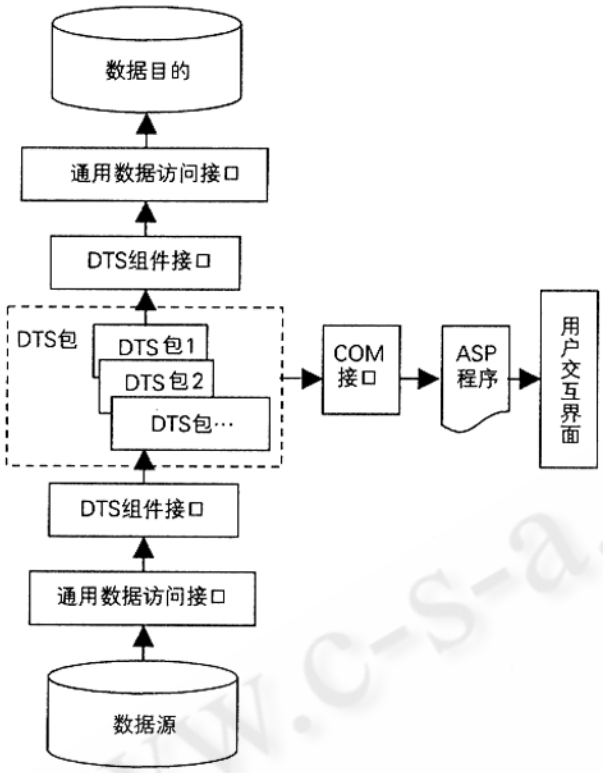

图 2　基于 DTS 组件的数据仓库中 ETL 解决方案设计框架

的总体架构和 ZMJ_DSS 系统划分的主题，我们设计了 18 个 DTS 包，具体 DTS 列表见表 2。

其中每一个包都由若干个数据抽取任务组成，每个数据抽取任务一般都是针对数据目的数据库中的某一个表的数据抽取。数据抽取任务由三部分组成：源数据表、目的数据表和抽取规则，其中的目的表一般是一个表，而源数据表则可能是一个或多个表中的数据，因此可以用 SQL 语句返回的查询结果表示源数据表，抽取规则一般是指在源数据表和目的数据表的基础上细化到字段的操作。下面将用一个列子说明 DTS 包的详细建立过程。

我们针对表 1 中的第 4 个 DTS 包—Data_Base_ODS，该包主要完成从 ZMJERP 到 ZMJODS 基础数据的装载。包中一个抽取任务是负责从源数据表 Dm_calendar 向目的数据表 D_time 中抽取数据，这个过程同时包含多个字段的抽取规则，其中大部分规则就是数据的直接复制，然也有一些规则是根据具体情况编写特定的程序脚本。如下便是将 Dm_calendar 表中的 week_day 字段抽取到 D_time 表中的 w_week 字段的相应规则处理脚本。

表 2　DTS 包列表

序号	DTS 包名称	描述
1.	Capmsplan_resource	从 Capmsplan 到 ZMJERP 数据装载—业务数据
2.	Hrerp_resource	从 Hrerp 到 ZMJERP 数据装载—业务数据
3.	Cbhs_resource	从 Cbhs 到 ZMJERP 数据装载—业务数据
4.	Data_Base_ODS	从 ZMJERP 到 ZMJODS 数据装载—基础数据
5.	Sales_ODS	从 ZMJERP 到 ZMJODS 数据装载—销售主题
6.	Finance_ODS	从 ZMJERP 到 ZMJODS 数据装载—财务主题
7.	Mps_ODS	从 ZMJERP 到 ZMJODS 数据装载—生产主题
8.	Inventory_ODS	从 ZMJERP 到 ZMJODS 数据装载—库存主题
9.	Scm_ODS	从 ZMJERP 到 ZMJODS 数据装载—采购主题
10.	Hr_ODS	从 ZMJERP 到 ZMJODS 数据装载—人力资源主题
11.	Base_Data_DSS	从 ZMJODS 到 ZMJDSS 数据装载—基础数据
12.	Sales_Base_DSS	从 ZMJODS 到 ZMJDSS 数据装载—销售主题(基础数据)
13.	Sales_Base_DSS	从 ZMJODS 到 ZMJDSS 数据装载—销售主题
14.	Finance_DSS	从 ZMJODS 到 ZMJDSS 数据装载—财务主题
15.	Mps_DSS	从 ZMJODS 到 ZMJDSS 数据装载—生产主题
16.	Inventory_DSS	从 ZMJODS 到 ZMJDSS 数据装载—库存主题
17.	Scm_DSS	从 ZMJODS 到 ZMJDSS 数据装载—采购主题
18.	Hr_DSS	从 ZMJODS 到 ZMJDSS 数据装载—人力资源主题

```
Function Main( )
  Select case DTSSource( " week_day" )
    case 1
        DTSDestination( " w_week" ) = " 星期一"
    case 2
        DTSDestination( " w_week" ) = " 星期二"
    case 3
        DTSDestination( " w_week" ) = " 星期
```

```
三"
        case 4
            DTSDestination ( " w_week" ) = " 星期
四"
        case 5
            DTSDestination ( " w_week" ) = " 星期
五"
        case 6
            DTSDestination ( " w_week" ) = " 星期
六"
        case 7
            DTSDestination ( " w_week" ) = " 星期
日"
    end select
      Main = DTSTransformStat_OK
    End Function
```

上面的抽取规则程序脚本主要完成如下功能：当 Dm_calendar 表中的 week_day 字段分别为 1、2、3、4、5、6、7 时，加载到 D_time 表中的 w_week 字段的数据分别对应为为星期一、星期二、星期三、星期四、星期五、星期六、星期日。项目中还有很多类似这样的抽取规则程序脚本，有的比这还要复杂，由于篇章的限制，这里只做一个简单的介绍，当然大多的规则的还是直接复制。

3.3 B/S 模式 ETL 程序

为了给系统管理员提供一个良好的用户界面以及方便他们使用，我们采用 ASP 和 COM 技术将后台 SQL SERVER 中的 DTS 包的执行和调度通过编程改成 B/S 结构，主要是通过 asp 访问存储在数据库系统 msdb 数据库中的 DTS 包元数据实现的。

数据转换服务（DTS）应用程序接口（API）是一组对象，用以封装帮助生成数据仓库的服务。可以在用支持自动化或 COM 的语言编写的应用程序中使用 DTS。下面就是利用 asp 访问 DTS COM 对象的部分脚本[5]。

```
<%
dim aa
aa = request. QueryString( " aa" ) '定义变量 aa，aa 的值通过 request 对象来传递
dim mypackage
    set mypackage = createobject ( " dts. package" ) 创建 dts. package 类对象实例 mypackage
    mypackage. LoadFromSQLServer " ZMJDSS " ," sa" ," zmjdss" ,0, , , ,aa
    mypackage. execute( ) '加载名为变量 aa 值的 dts 包并执行
    iStatus = True
    response. Write " <table width =333 border =0 align = center >"
    For i = 1 To mypackage. Steps. Count 访问包的每个步骤，提取其状态信息
    If mypackage. Steps( i) . ExecutionResult = 1 Then
        iStatus = False
        response. Write ( " < tr > < td > 步骤:" + mypackage. Steps( i) . description + " 执行失败" + " < /td > < /tr > " )
    else
        response. Write ( " < tr > < td > 步骤:" + mypackage. Steps( i) . description + " 执行成功" + " < /td > < /tr > " )
    End If
    Next
    response. Write ( " < tr > < td > < /tr > < /td > " )
    response. Write ( " < tr > < td > < /tr > < /td > " )
    If iStatus = True Then '若包中有一个步骤执行失败，则该包执行失败
    response. Write ( " < tr > < td > " + mypackage. Description + " 执行成功 < /td > < /tr > " )
    else
    response. Write ( " < tr > < td > " + mypackage. Description + " 执行失败 < /td > < /tr > " )
    End If
    response. Write " < /table > "
    Set mypackage = Nothing
    % >
```

由于限制篇幅的关系，我们省略用 ASP 访问和执行调度 dts 包存储过程的程序以及其他 ZMJ_DSS 项目中的 ETL 子系统中所包含程序，只列出了最核心的程

序代码,其他的部分读者可以参照上面或查阅相关资料自己研究。

4 结束语

ETL 是建立数据仓库系统的最重要环节之一,在一个数据仓库项目中,约 80% 的工作量都花费在此[6]。本文以张家口煤机厂决策支持系统项目为背景,应用微软 DTS 组件和 ASP 技术实现该项目的 ETL 子系统,同时也提出了一种 B/S 模式的数据仓库中的 ETL 方案。实践证明,该解决方案能够方便 ETL 管理员操作,打破了以往 ETL 管理员只能在服务器上执行任务的局限,具有很好的实用价值。但该方法的通用性和执行效率问题还有待于改进和提高。

参考文献

1 王珊,数据仓库技术与联机分析处理[M],北京:科学技术出版社,1999.

2 张宁、贾自艳、史忠植,数据仓库中 ETL 技术的研究[J],计算机工程与应用,2002(24):213 -215.

3 辽宁工程技术大学系统工程研究所.张家口煤矿机械有限公司决策支持系统概要设计说明书[R],2004.

4 汤晓兵、汪美霞,SQL Server 2000 与异构数据源之间的传输转换技术[J],山东建筑工程学院报,2004,19(2):52 -54.

5 辽宁工程技术大学系统工程研究所.张家口煤机厂决策支持系统详细设计说明书[R],2004.

6 武彦峰、朱仲英,基于 DTS 组件的数据仓库的数据抽取工具的设计与实现[J],微型电脑应用,2004,20(2):1-4.