面向业务的数据集成系统设计与实现①

时俊苓 叶 丹 (中国科学院 研究生院 北京 100049;中国科学院 软件研究所 北京 100190)

摘 要: 在企业数据集成过程中，大多采用适配器实现异构数据源的访问，针对每个数据源开发一个特定的数据源适配器，存在开发难度大，开发周期长的问题。同时，由于各个业务系统数据的复杂性，导致数据集成系统的配置、部署复杂。自主研发的面向业务的数据集成系统，不仅解决了分布式环境下异构数据的集成，同时，使数据集成系统具有良好的扩展性及部署的简单性。介绍了系统的体系结构，给出了数据源适配器开发框架，提出了一种基于中间表的增量数据获取及发送方法，并结合某百货集团业务需求，给出了数据集成的方案。

关键词: 数据集成；中间表；数据增量；数据发送；数据源适配器

Design and Implementation of A Business-Oriented Data Integration System

SHI Jun-Ling, YE Dan

(Graduate School of Chinese Academy of Sciences, Beijing 100049, China;
Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: During the enterprise data integration, the heterogeneous data source is mostly accessed through adapters. It is difficult and long to develop a specific source adapter. At the same time, configuration and deployment of data integration system are complex due to the complex business system. This paper develops a business-oriented data integration system, which cannot only resolve the data integration between the heterogeneous source in Distributed Computing Environment，but also make the system's deployment and expansibility simple. It introduces the system's architecture, gives a development framework of source adapter and a method to data capture and sending. As a application case, a resoluble approach about some general merchandise group is also put forward.

Keywords: data integration; middle table; data incremention; data sending; data source adapter

1 引言

随着企业信息化建设的发展，企业建立了众多的信息系统，以帮助企业进行内外部业务的管理。但是，企业各系统的数据是分布的、异构的， 为了共享这些业务数据，需要一个数据集成系统来完成数据的共享与转换。

在企业数据集成过程中，使用适配器实现各个数据源的连接访问是主要方法，但是，由于数据源的多样性、异构性，使适配器也变得多样，复杂。对每种数据源都重新开发一个适配器，既增加了开发难度，也增加了开发工作量。因此，有必要针对典型的集成模式，提供一个通用的适配器开发框架，将可复用的复杂功能封装起来，留出扩展接口。当开发一个新适配器时，只实现扩展接口。这样提高了开发效率。另外，由于各个业务系统数据的复杂性，导致数据集成系统的配置、部署复杂。

为解决这些问题，我们开发了面向业务的数据集成系统，以适配器的方式，实现对数据的访问。针对

① 基金项目:国家高技术研究发展计划(863)(2007AA01Z149,2007AA04Z148)
收稿时间:2009-10-19

数据库提供了适配器的开发框架。在具体的数据库业务数据访问时，采用中间表的方式实现基于变量的增量数据的获取和发送。 既满足了基本的数据集成要求，即数据共享与交换，同时具有灵活的扩展性与部署的简单性。

2 数据增量获取方法

目前，常用的数据增量获取方法有：触发器、时间戳、快照、日志对比[1]。

① 触发器：在用户表上要建立插入、修改、删除三个触发器，每当源表中的数据发生变化，就被相应的触发器将变化的数据写入一个临时表，抽取线程从临时表中抽取数据，临时表中抽取过的数据被标记或删除。

② 时间戳：在源表上增加一个时间戳字段，当进行数据抽取时，通过比较系统时间与时间戳字段的值来决定抽取哪些数据[2]。

③ 快照：数据的快照是指数据的一个备份。在上次发送时保留当时的快照，当前发送时，通过比较当前数据与上次发送时快照，得到增加、删除和修改的数据。

④ 日志对比：通过分析数据库自身的日志来判断变化的数据。

触发器方式，需要在业务系统数据库中创建触发器，不仅要求对数据库的权限较大，也会影响业务系统性能。.快照方式，需要对发送前后的纪录逐条比较，性能较差。时间戳要修改业务系统的数据库设计，当业务系统设计实现完以后，修改数据库设计，对系统的修改很大。日至对比，需要对数据库系统日志很了解，不同数据库日志不同，分析日志也很复杂。

我们提出一种，基于中间表的增量获取方法，该方法可以有效地实现复杂业务数据的增量数据交换，同时通过有效获取中间表中业务信息，使得数据交换具有更多的业务相关性。同已有的快照、触发器相比，中间表的增量方式在处理复杂业务表时具有更大的优势。基于中间表增量方式的数据获取具有如下优点：

第一、 由于有效地将触发器的管理转移到业务系统，中间表增量方式可实现多表的数据增量，这是单一触发器方式难以实现的。

第二、 由于中间表是基于业务系统建立的，其管理和维护由业务系统完成，因此只需要赋予数据集成系统对中间表的读、写权限就可以完成数据交换功能，从而避免了权限太大带来的安全问题。而快照和触发器方式往往需要赋予表和触发器的创建、删除和修改权限。

第三、 由于是基于有限数目中间表的条件查询，因此同快照方式相比，速度有很大提升。

3 系统体系结构

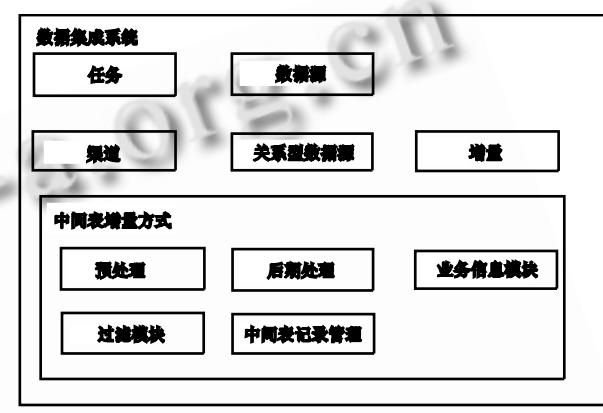

图 1　数据集成系统体系结构

(1) 任务模块：数据的提取、转换、传输和加载的全过程，通过任务定义的方式定义数据处理流程，完成在各种网络条件下的数据复制和传送操作。任务可以手工执行和自动执行，特别是可以灵活定义触发条件，如确定时刻触发、确定间隔触发，实现数据的自动复制；同样，也可预先设定任务停止触发的时间。发送端任务调用数据源进行数据读取，调用渠道进行数据的发送。接收端，渠道自动接收数据，任务调用数据源进行数据的保存。

(2) 数据源模块：完成数据的读取、转换、加载，支持关系型数据，如：数据库。本文重点讨论关系型数据库中业务数据增量复制实现方法。

(3) 渠道模块：基于消息队列，完成数据发送，接收。通过接收方 IP，创建发送渠道，接收渠道。

(4) 中间表增量方式模块包括如下几个模块：

① 预处理模块：调用中间表管理模块，获取中间表记录,调用过滤模块，将查询语句中的变量进行过滤赋值。

② 过滤模块:读取数据前，对数据进行过滤。完成查询语句中过滤条件的赋值。

③ 中间表管理模块主要功能：

获取指定数目的主中间表记录，删除主中间表记录；获取当前操作的主、次中间表记录；管理次中间表的记录。

④ 后期处理模块：完成提取数据后对中间表数据的删除，对业务信息模块的维护。

⑤ 业务信息模块：将业务数据写到中间表。

4 关键技术实现

4.1 基于变量的数据获取

在业务系统数据库中，创建中间表，配置一些读取数据的条件，若有多个目的接收方，中间表中配置接收方 ID，在读取数据，发送数据时，根据中间表中的读取条件构造查询语句，根据接收方 ID，选择发送渠道。

中间表分为主中间表和次中间表，主中间表对应业务表中的主表，次中间表对应业务表中的次表。业务系统根据需要将记录保存到主次中间表中。

在用于交换的数据查询 sql 语句中加入条件变量，在提取数据前，读中间表的记录，根据中间表中的记录值对变量进行赋值，然后进行数据提取；若提取成功则将中间表记录删除，循环获取下一条中间表记录。

处理流程如下：

① 获取中间表中一条满足条件的记录，

② 根据①的结果，为业务表查询 sql 语句中的变量赋值，若有次中间表则循环①、②。

③ 将设有变量的 sql，作为发送语句，进行数据提取、发送操作。

④ 若③成功，则将主中间表记录删除，循环①，直到主中间表没有满足的记录；否则退出。

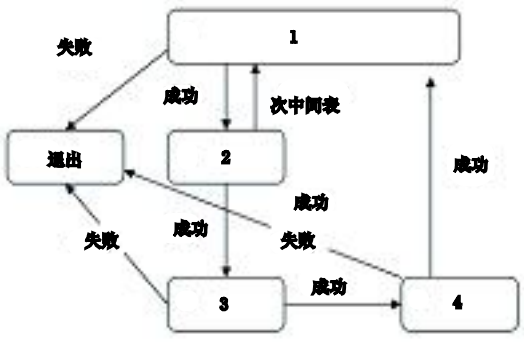

图 2　中间表增量方式处理流程

4.2 基于变量的数据发送

数据交换本质而言是基于关联的。基于关联配置的数据交换的特点是：一次配置多次交换。在接收方数据结构相同情况下，如果进行多次一对一的关联，显然是烦琐、重复的。因此进行一次关联配置，通过简单修改目的地，实现一对多的数据交换。

发送渠道以两种方式对数据进行传输：一种是发布模式，在不知道目的端 IP 的情况，由发送端把数据发到指定目录，接收方到该目录下取数据。另一种是订阅模式，接收端在发送端注册 IP，发送端根据 IP，把数据发送给接收端。

采用基于变量的数据发送方法，在中间表中设置发送的目的地，读取数据后，根据中间表中的目的地，动态选择发送渠道，通过发送渠道发送数据。发送渠道如果指定了目的 IP，则以订阅模式将数据发送给接收方，发送渠道不知道目的 IP，则以发布模式将数据发送到指定目录，由接收方自动获取。

4.3 任务的调度执行

任务的调度需要完成两方面的功能，分别是定期启动自动任务和按照触发关系启动任务。任务的调度模块有一个专门的调度线程。该线程启动一个任务后，会根据下次启动时间的顺序对所有自动任务进行排序，然后等待最近任务执行时刻的到来，到了该时刻后，启动该任务，依次循环。

调度算法：

```
While(true)
 {
  If (自动执行任务链表是空)
   {
     线程挂起;
   }
  获取第一个任务的下次执行时间;
  If (当前系统时间 < 任务的下次执行时间)
  {
   Sleep(任务的下次执行时间 - 当前系统时间);
   Continue;
  }
  Else
  {
  执行该任务，并从自动任务执行链表中出;
  重新计算该任务的下次执行时间， 通过排序，将该
```

任务插入到链表中。

}

}

4.4 数据源适配器开发框架

针对主流的数据库，采用 odbc 技术，实现统一的访问接口，如：数据库的连接，数据库数据的读写等，将这些共有的功能封装为静态库。当开发一个新的数据库适配器时，引用该静态库，只实现数据库特殊的接口。特殊的接口包括：主外键的获得、主键冲突错误的判断、ODBC 数据类型的匹配等等。这些接口往往通过 ODBC 接口获得的结果是不同的，需要在各自的适配器中，单独实现。开发完的适配器编译成动态库的形势，在服务器端动态加载和卸载。

5 系统的应用

某百货集团，集团总部数据库中，存放有各个子公司的业务数据，如：销售信息、订单信息、库存信息等。各个供应商在自己本地的查询系统中，查询自己所供应商品的销售，库存等信息。因此，需要百货集团总部定期的将各个供应商需要的信息，发送到供应商本地。图 3 是一个一对多数据发送的配置模式。

发送配置：定义查询语句，该查询语句的查询条件是基于中间表中的存放的变量。

发送渠道：定义接收端的 IP，增加一个接收端，需要在发送端，增加一个对应的发送渠道。接收端接收配置：定义接收目的表，目的字段，对发送数据进行模式匹配。

接收渠道：定义接收端数据的接收目录。发送端根据接收端 IP，将数据发送到该目录。

该部署模式满足某百货集团数据集成的需求。集团总部作为发送端，各个供应商作为接收端。由于各个供应商的查询系统，数据库相同，在数据集成系统部署时，把供应商作为查询数据，发送数据的变量，只完成一次配置。当增加一个供应商时，只需要在集团总部，增加一个指向该供应商的发送渠道，在发送数据时，根据供应商这个变量，选择该发送渠道。在供应商接收端，创建具体的接收渠道，完成到不同供应商一对多的数据发送。

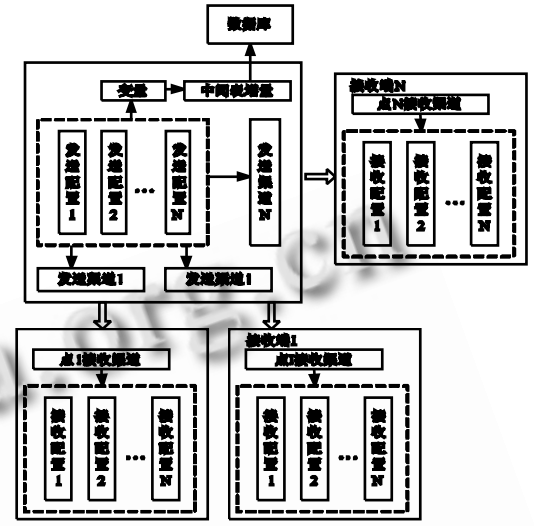

图 3　一对多数据的发送

6 结语

我们开发出的面向业务的数据集成系统，已经在某百货集团广泛应用，该系统的特点：

① 业务相关性

含有业务属性的数据交换、具有业务含义的日志管理。

② 可扩展

新增业务的可扩展性

③ 部署简单

软件部署的简单性

④ 支持大规模使用

大规模使用的可靠性

参考文献

1 章水鑫,徐宏炳,于立.增量式 ETL 工具的研究与实现.现代计算机, 2005,3:6－10.

2 夏榆滨,黄善琦.基于 ETL 技术的企业信息集成研究.微计算机信息, 2006,22(10-3):190－192.