文章编号:1006-2475(2009)08-0101-04

基于 SSIS 平台的 ETL 设计与实现

朱丽雅[1],曾 成[2],向 青[3]

(1. 四川行政学院计算机科学与工程教研部,四川 成都 610000;
2. 中国人寿四川省分公司信息技术部,四川 成都 610000;3. 四川行政学院学历教育部,四川 成都 610000)

摘要:通过中国人寿四川统计信息系统的设计与实现,本文探讨通过 SSIS 系统平台实现 ETL 解决方案,如何针对来自不同应用系统、不同数据平台、不同数据源形式的源数据系统存在的数据质量的差异性、缺乏一致性等问题,将数据从源数据系统中抽取、转换成数据仓库需要的格式和统一数据类型,并正确加载到数据仓库中,为统计分析系统的实现提供高质量的基础数据。

关键词:数据仓库; ODS; SSIS; ETL
中图分类号:TP311.131 文献标识码:A doi:10.3969/j.issn.1006-2475.2009.08.028

Design and Implementation of ETL Base on SSIS

ZHU Li-ya[1], ZENG Cheng[2], XIANG Qin[3]

(1. Department of Teaching Laboratory on Computer Science and Engineering, Sichuan Administration College, Chengdu 610000, China;
2. Information Technology Department, Chinalife Sichuan Branch, Chengdu 610000, China;
3. Academic Education Department, Sichuan Administration College, Chengdu 610000, China)

Abstract: Base on design and implementation of the ChinaLife Sichuan statistical information system, this paper discusses how to design a good ETL base on SSIS solution to solve the difference of data quantity and the lacking of consistency of the database system, which come from different application system and different of data platform. The paper also discusses how to load the data from different application system and transform the data type that the target system want to and load the data to the data warehouse, at last the ETL process provides the data of the high quantity for statistical information system.

Key words: data warehouse; ODS; SSIS; ETL

0 引 言

建设数据仓库需要集成来自多种业务数据源中的数据,这些数据源可能处在不同的硬件和操作系统之上,在编码、命名、数据类型、语义等方面都存在较大的冲突,因此如何向数据仓库中加载这些数量大、种类多的数据,成为建立数据仓库所面临的一个关键问题。如果最终加载的信息不准确,那么这个数据仓库便会形同虚设,所以将操作数据导入数据仓库的过程,必须经过精心的规划和设计,并建立一个相对独立的系统来完成数据转换工作。ETL 解决方案是其成功与否的关键[1]。

本文通过中国人寿四川统计信息系统的设计与实现,探讨通过 SSIS 系统平台实现 ETL 解决方案,解决实现数据仓库 ETL 过程出现的问题,如:各业务系统的复杂性,各个源数据系统的数据结构、格式、类型、定义各不相同,各源系统的数据质量的差异性,源系统的数据缺乏一致性等,为数据仓库系统的实现奠定坚实的基础[2]。

1 相关技术

1.1 ODS(Operational DataStore,操作型数据存储)[3]

ODS 是用于支持企业日常全局应用的数据集合。ODS 作为一个中间层次,一方面,它包含企业全

收稿日期:2009-05-22
作者简介:朱丽雅(1974-),女,甘肃酒泉人,四川行政学院计算机科学与工程教研部讲师,硕士,研究方向:数据库技术;曾成(1980-),男,四川成都人,中国人寿四川省分公司信息技术部工程师,研究方向:数据库技术;向青(1973-),女,四川成都人,四川行政学院学历教育部助理工程师,研究方向:计算机基础教育。

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；另一方面，它又是一种面向主题的、集成的数据环境，且数据量较小，适合于辅助企业完成日常决策的数据分析处理。因此，保存在 ODS 中的数据具有 4 个基本特点：面向主题的、集成的、可变的、数据是当前或接近当前的。

1.2 ETL 解决方案[4]

包括数据抽取(Extract)、转换(Transform)与清洗、数据加载(Load)与调度，ETL 系统将贯穿整个数据仓库系统的全过程[5]。ETL 要完成的工作就是在数据仓库和业务系统之间搭建起一座桥梁，确保新的业务数据源源不断地进入数据仓库。在进行 ETL 的过程中，往往要面对大量的业务逻辑和异构环境，因此 ETL 的主要作用在于屏蔽复杂的业务逻辑，从而为各种基于数据仓库的分析和应用提供了统一的数据接口[6]，这也是构建数据仓库最重要的意义所在。这个过程还关系到数据的质量，所以是数据仓库应用的基石。

1.3 ETL 方式选择

在选择 ETL 工具时，可以从 ETL 对平台的支持、对数据源的支持、数据转换功能、管理和调度功能、集成和开放性、对元数据管理等功能出发进行考虑[7]。

数据仓库项目在进行 ETL 开发时，一般有两种方式可供选择，即 ETL 工具和手工编码。目前市场上主流的 ETL 工具可以分为两类[8]：一类是专业的 ETL 厂商的产品，如 Ascential Datastage、Informatica、Sagent Solution 等。这类产品一般都有较完善的体系结构和久经考验的产品线，产品提供的功能非常复杂和详尽，但产品价格也非常高昂；另一类是整体数据仓库方案供应商，他们在提供数据仓库存储、设计、展现工具的同时也提供相应的 ETL 工具，如 Oracle Warehouse Builder、IBM Warehouse Manager、SQL Server 2005 Integration Services (SSIS) 等。这类产品一般对自己厂商的相关产品有很好的支持，因此能够发挥最大的效率，但结构相对封闭，对其他厂商产品的支持也有限[9]。

2 ETL 在保险统计信息系统中的应用

中国人寿目前使用的数据服务平台，通过在各个业务系统数据库(包括 CBPS8、CBPS7、年金险系统、万能险系统、统括业务系统、AMIS 代理人管理系统等)上建立 Trigger，将这些数据库中大部分实时更新的数据，刷新到数据服务平台中。它较好地完成将异构的、跨平台的业务系统数据或者外部数据经过清洗、转换后存放到统一的数据平台里这一过程。但该平台实际是一个操纵型数据存储，其包含了最小粒度、最详细的事务级的细节数据，因此其对业务分析和决策支持应用支持较差。介于这种情况，提出了以数据服务平台和各业务系统物理镜像数据库为基础来源，建立面向全省的业务分析和决策支持系统这一方案，为全省各级分公司的业务分析和决策提供服务。

2.1 人寿统计信息系统体系架构

引入 ODS 层，形成 DB-ODS-DW 三层结构[10]，即通过在各省公司建立 ODS 层，完成对地市细节级数据的抽取、转换和存储工作，各省公司在 ODS 的基础上构建数据仓库，形成 DB-ODS-DW 三层结构。同时在总公司也建立 ODS 层，从省公司 ODS 层提取数据，并作为总公司数据仓库的数据源[11]。这样就形成了省公司-总公司两层统计信息平台，系统整体架构如图 1 所示。

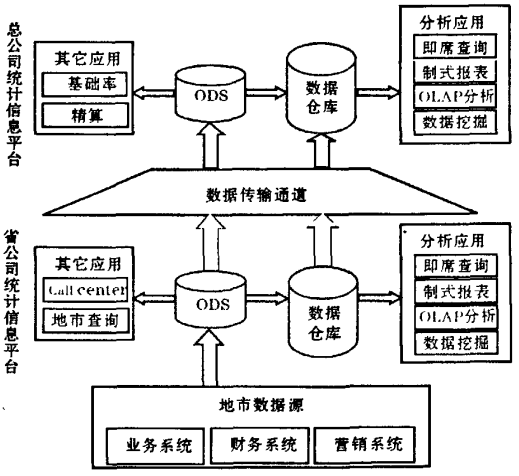

图 1 基于 DB-ODS-DW 的系统架构

在这样的系统架构下，原来全部存储在总公司的地市级细节数据被分散到了各省公司的 ODS 层中，极大地缓解了总公司的存储压力。各省公司分别完成地市级细节数据到省公司 ODS 层的数据采集和清洗转换，最后再统一集中到总公司 ODS 层，这种分布式集中的方式使得数据集中效率大大提高。另外，省公司 ODS 层可以支持需要当前细节数据查询的应用(如 Call Center 应用)，以及把原来建立在各个地市应用系统上的查询部分转移到省公司 ODS 层，减少对各地市应用系统的压力。同时，省公司 ODS 层还可以为省公司数据仓库和总公司 ODS 层提供数据，使总公司建立基于 ODS 的数据仓库应用和其他关键应用(如精算、基础率等)，使各省公司建立自己的数据仓库应用，从而充分提高了系统利用率。

2.2 本项目 ETL 策略

统计信息系统数据仓库是以数据服务平台为基础

的面向主题的数据集市,数据服务平台已完成了将异构、跨平台的业务系统数据或者外部数据经过清洗、转换、整合存储到其数据库中这一过程,但由于数据服务平台是 ODS 层,其保存的都是实体关系的明细数据,所以统计信息系统数据仓库(ScQuery)将着重完成将数据服务平台的数据转换和整合到 ScQuery 数据库中这一过程,其完成的是从 ODS 向数据集市的抽取[12]。

抽取转换总体思路:

(1)为了提高抽取速度,对源表进行直接抽取,不对任何表进行关联,不作代理键转换,只生成度量值,然后将数据直接插入到临时表中,但数据中要包括转换的自然关键字,同时将代理键字段设置为默认值 0;

(2)对由于自然关键字为空而引起无法生成代理键的情况进行特殊处理;

(3)对临时表中的事实数据进行代理键的生成;

(4)临时表中的事实数据插入事实表中。

本文选用保费收入主题进行 ETL 过程的分析设计:

表 1 为保费收入主题从统计服务平台到统计信息数据仓库(ScQuery)的纪录系统定义、转换。

表 1　从统计服务平台到统计信息数据仓库的记录系统定义

源表	源字段	目标表	目标字段	转换描述
B01 实收费流水表、B02 实付费流水表	Mio_date(Datetime)	Factpremium 事实表-保费收入	Miodatekey(int)	通过维度表 dimtime 关联转化
	In_force_date(datatime)		Validatekey(int)	通过维度表 dimtime 关联转化
	Mgr_branch_no(varchar(6))		Mgrbranchkey(int)	通过维度表 dimbranch 关联转化
	Pol_code(varchar(6))		Polkey(int)	通过维度表 dimpol 关联转化
	Moneyin_itrvl(varchar(1)) Moneyin_dur(int)		Payitrvlkey(int)	通过维度表 dimpayitrvl 关联转化
	Sales_channel(varchar(2)) Sales_branch_no(varchar(6)) Sales_no(varchar(8))		Salesclerkkey(int)	通过维度表 dimsalesclerk 关联转化
	Mio_log_id(varchar(30))		Miold(varchar(30))	直接抽取获得
	Cntr_id(varchar(30))		Cntrid(varchar(30))	直接抽取获得
			Newpremium(decimal(18,2))	判断提取
			Standardpremium(decimal(18,2))	存储过程计算获得
			Continuepremium(decimal(18,2))	判断提取
	Amt(decimal(18,2))		Totalpremium(decimal(18,2))	直接抽取获得

抽取策略的说明:

(1)由于保费收入统计是指本年发生的保单收入情况,因此只抽取本年产生的收费记录。根据统计,中国人寿四川公司一年产生的收付费记录大概是一千万条,数据量较少,抽取一次存量大概在三十分钟左右完成,因此在这里选用了存量抽取的策略。

(2)由于抽取的是财务实收费流水表,因此表中包括了较多的冲正记录,这部分数据对统计分析来说是没有意义的,因此在这里直接做了清洗。

(3)在对脏数据进行处理时,需找出具体原因,并听取业务部门的意见,针对不同的情况,制定修正策略,以保证数据最接近真实值。

2.3 SSIS 实现 ETL

SSIS 是生成高性能数据集成和工作流解决方案(包括数据仓库的提取、转换和加载(ETL)的平台[10])。

SSIS 实现保费收入 ETL 过程说明:

首先清洗各维度表和事实表 factpremium,接着对各维度表和事实数据进行抽取,需要生成的较重要的维度表一共 5 张,但它们的抽取策略是不相同的。dimtime 维度表是通过脚本任务生成的,dimpayitrval 是通过从文件抽取数据生成,这是因为它的数据较少,是通过手工定义的,而其它 3 张维度表则是从数据服务平台的 6 张字典表中抽取生成,它们分别是 dimpol、dimsalesclerk 和 dimbranch。在抽取维度表的同时,为了提高抽取速度,定制了并行抽取的策略,同时对事实数据进行了抽取。首先将事实数据抽取到 tmp_factpremium 表中,即包括实收费事实也包括实付费事实,这两部分数据分两次分别从数据服务平台的 2 张表中进行抽取,在抽取完成后,就对这部分事实数据进行数据清洗,包括 null 的设定等。在所有维度表都已生成、事实数据也清洗完成后,再将事实数据从 tmp_factpremium 表中插入到事实表 factpremium 中,这时保费收入主题的 ETL 抽取也就完成了。

把 Microsoft SQL Server 2005 Integration Services (SSIS)作为 ETL 开发平台,能较方便地实现整个 ETL 开发。系统通过图形化用户界面来设计实现整个 ETL 过程,十分直观,开发速度也非常快。最后通过 SSIS 包的部署和调度,最终实现 ODS 系统数据的 ETL 任务。

3　结束语

数据是企业进行任何事务的前提,ETL 的目的正是提供综合且高品质的数据[13],因此它必然成为企业各类应用的基础,为众多的高层信息系统提供服务。具备良好的通用性是未来数据 ETL 软件占领市场的必要条件,这就要求它:支持尽可能多的数据库管理系统(DBMS)、文件系统和数据采集、处理系统;能够跨网络、跨平台使用;具备良好的可扩展性,对于新的应用能够以较小的代价、通过预定的应用程

序接口(API)或标准化语言接口编程实现互联[14]。相关技术的发展，如元数据的标准化、程序逻辑与数据的统一化，为提高 ETL 通用性提供了动力。

参考文献：

[1] 张宁,贾自艳,等. 数据仓库中 ETL 技术的研究[J]. 计算机工程与应用,2002,38(24):213-216.

[2] 陈弦,陈松乔. 基于数据仓库的通用 ETL 工具的设计与实现[J]. 计算机应用与研究,2004,21(8):214-216.

[3] Inmon W H. Building the Operational Data Store[M]. New York:John Wiley & Sons,Inc.,1996.

[4] Inmon W H. 数据仓库[M]. 北京:机械工业出版社,2002.

[5] Ralph Kimball,Joe Caserta. The Data Warehouse ETL Toolkit:Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data[M]. Indianapolis: Wiley Publishing, Inc.,2004.

[6] 连立贵,金凤,蔡家楣. 数据仓库中的数据提取[J]. 计算机工程,2001,27(9):61-62,99.

[7] 周宏广,周继承,彭银桥,等. 数据 ETL 工具通用框架设计[J]. 计算机应用,2003,23(12):96-98.

[8] 吴悦. ETL 工具点评[J/OL]. http://media.ccidnet.com/media/ciw/1214/d1301.htm,2003-05-12.

[9] 苏鹏,李钊,王文,等. 基于 SSIS 企业数据集成系统的技术实现[J]. 计算机应用与软件,2008,25(9):179-180,202.

[10] 王霓虹,刘美玲. ODS 数据仓库新技术的研究与应用[J]. 信息技术,2004,28(11):8-11,31.

[11] 王兴渝. 基于 ODS 的数据仓库技术在人寿统计信息系统中的应用研究[D]. 北京:清华大学硕士学位论文,2006.

[12] 郑擎宇,郭研,左春. 保险业数据参考模型对 ETL 的影响和作用[J]. 计算机系统应用,2007(3):50-54.

[13] 张蓓,赵莉. 浅谈数据仓库中 ETL 的重要性[J]. 科技信息,2008(18):82,64.

[14] 刘强,翁惠玉. 基于电信行业的 ETL 系统的设计与实现[J]. 计算机工程,2004,30(z1):30-31,42.

(上接第 100 页)

(3) 生成目标 XML 文档：获得 XML 文档的内容后，先将执行结果嵌入到元素树中，然后再将元素树序列化加上 XML 文档序列，即可得到目标 XML 文档。在将执行结果嵌入到元素树的时候，需要对执行结果作一定的合法性检查。例如对于上述的 father 元素，其执行结果中 son2 与 son3 是不能够在同一条记录中同时出现的。若同时出现了，则提示用户进行选择或直接报错。

3 结束语

本文着重介绍了 XML 在多层体系结构中的数据传输的一些关键技术和实现细节。首先给出了基于 XML 的多层体系结构的模型，然后针对模型中的前两层进行了详细的介绍：在客户端发送请求并接受来自服务器的返回结果，利用 XML 数据岛技术返回数据给客户端以各种要求的形式显示；在服务器端，引入数据转换模式，进行 XML 和关系数据之间的数据转换，以达到异构数据库之间的数据交换目的。

参考文献：

[1] 叶枝平,李振坤,刘竹松,等. 基于 XML 的数据交换平台的研究与设计[J]. 微计算机信息, 2008,24(9):243-244,229.

[2] 张福军. 基于 JAVA 与 XML 技术的数据交换工具的设计[J]. 科技创新导报, 2008(9):22-23.

[3] 蔡振闹. 数字化校园数据交换平台的研究与设计[J]. 福建电脑,2008,24(5):154-155.

[4] 李文翔,肖军模. XML 技术在数据交换中的应用[J]. 电脑知识与技术, 2008,2(17):1543-1545.

[5] 吴问平,刘锋. 基于关系数据库的 XML 数据交换技术[J]. 铜陵学院学报, 2008(3):65,70.

[6] 金蓓弘,刘志军. 数据传输工具 DataTrans 的设计与实现[J]. 计算机工程与应用,2001,37(17):7-11.

[7] 李钊,曹亮,唐春华,等. 一个 XML 的数据模型及其存储策略[J]. 计算机应用研究,2001,18(11):139-141.

[8] 谷长勇,徐志伟,褚兴军. XML 结构和关系数据库的一种形式化映射[J]. 计算机工程,2001,27(11):16-17.

[9] 王海波. 基于 XML 的数据交换的实现[J]. 计算机应用,2006,21(4):67-68.

[10] 李雯,谢辅雯,邹道明. XML 数据交换技术的应用与研究[J]. 计算机与现代化,2008(1):17-19.

[11] 方美琪. XML 及其在电子商务中的应用[M]. 北京:清华大学出版社,2003:56-59.

[12] 张红梅,梁允荣. 基于 XML 实现电子商务平台的分析与研究[J]. 计算技术与自动化,2005,24(2):115-117.

基于SSIS平台的ETL设计与实现

作者： 朱丽雅， 曾成， 向青， ZHU Li-ya， ZENG Cheng， XIANG Qin

作者单位： 朱丽雅,ZHU Li-ya(四川行政学院计算机科学与工程教研部,四川,成都,610000)， 曾成,ZENG Cheng(中国人寿四川省分公司信息技术部,四川,成都,610000)， 向青,XIANG Qin(四川行政学院学历教育部,四川,成都,610000)

刊名： 计算机与现代化 ISTIC

英文刊名： COMPUTER AND MODERNIZATION

年，卷(期)： 2009(8)

参考文献(14条)

1. 刘强;翁惠玉 基于电信行业的ETL系统的设计与实现[期刊论文]-计算机工程 2004(z1)

2. 张蓓;赵莉 浅谈数据仓库中ETL的重要性[期刊论文]-科技信息 2008(18)

3. 苏鹏;李钊;王文 基于SSIS企业数据集成系统的技术实现[期刊论文]-计算机应用与软件 2008(09)

4. 吴悦 ETL工具点评 2003

5. 周宏广;周继承;彭银桥 数据 ETL 工具通用框架设计[期刊论文]-计算机应用 2003(12)

6. 连立贵;金凤;蔡家楣 数据仓库中的数据提取[期刊论文]-计算机工程 2001(09)

7. Ralph Kimball;Joe Caserta The Data Warehouse ETL Toolkit:Practical Techniques for Extracting,Cleaning,Conforming,and Delivering Data 2004

8. Inmon W H 数据仓库 2002

9. Inmon W H Building the Operational Data Store 1996

10. 陈弦;陈松乔 基于数据仓库的通用 ETL 工具的设计与实现[期刊论文]-计算机应用与研究 2004(08)

11. 郑擎宇;郭研;左春 保险业数据参考模型对ETL的影响和作用[期刊论文]-计算机系统应用 2007(03)

12. 王兴渝 基于ODS的数据仓库技术在人寿统计信息系统中的应用研究 2006

13. 王霓虹;刘美玲 ODS 数据仓库新技术的研究与应用[期刊论文]-信息技术 2004(11)

14. 张宁;贾自艳 数据仓库中 ETL 技术的研究[期刊论文]-计算机工程与应用 2002(24)