保险业数据参考模型对 ETL 的影响和作用①

Insurance Data Reference Model's Influence and Effect on ETL

郑擎宇 郭 妍 左 春 (中国科学院研究生院 北京)

摘要:本文在简略回顾保险业通用数据参考模型之后,以数据仓库技术中的 ETL 过程为着眼点,对基于保险业数据参考模型的 ETL 过程进行设计和分析,从而有针对性地提出一种保险行业数据抽取转换和加载的方法,最后探讨该设计在保险公司的应用。

关键词:数据参考模型 数据仓库 ETL 数据抽取转换加载 保险

1 引言

保险行业在中国的建立和发展经历了半个世纪的时间,伴随这半个世纪计算机技术飞速的发展,保险行业纷纷启动了计算机信息系统的建设,将保险公司经营模式从效率低下的手工方式逐渐转变为电子化方式,并且每年投入 IT 系统建设的资金不断增多,综合 IDC、CCID、麦肯锡在 2001 年的市场调查数据可知,中国保险公司对 IT 服务部分的资金投入占其 IT 系统总投入的 6%,在保险应用软件的投入方面占 7%。保险应用软件在不同时期不同规模的保险公司内的发展也是千差万别,作为中国最早的保险公司——人保财险、国寿,其保险应用软件也从庞大而分散的形式向集中化迈进,在这个过程中,面向操作型的核心业务系统已经满足不了市场化保险公司迫切的需要——从电子化的数据中分析经营不足、获得市场信息、预测发展,另一方面,保险公司经过十多年的计算机信息系统的建设已经积累海量的数据,那么,怎样将这些数据进行有效地提炼、发现信息、挖掘知识,将成为各家保险公司新的竞争工具。

数据仓库的出现和成熟为保险业解决上述问题提供了技术基础,无论在怎样的解决方案中,将正在使用的、遗留的、外部的系统所产生的数据进行抽取,并根据合理的规则进行清洗转化、高效率地加载到系统当中,整个过程都会是至关重要的,本文试图就保险业的 ETL 过程进行分析和设计,并说明保险业数据参考模型对 ETL 的影响和作用。

2 保险业通用数据参考模型回顾

在[1]中,技术工作者对保险业通用数据参考模型进行了细致的说明,描述了保险业通用数据参考模型的原理,如图 1 所示,该模型使用"纵向—关系、横向—实体"的分类方法将事物分成两类,纵向构成中的事物是由关系表及关系表的关系组成,关系表中定义了事物的属性,这些属性自由地描述了事物的特征,而在这些属性中自由度相对固定的部分形成了维;横向构成是可以被纵向构成中各个事物所使用的非标量域的通用代码字典,或者说是维的定义。

以保险业通用数据模型为指导可以快速地完成操作型源系统的数据库设计,遵循相同的数据模型,在纵向构成中根据保险业务分析的需要按照维方向进行累计汇总、滚动汇总、派生、演化等数据处理并加以存储,于是可以得到分析型的系统,并保持了两类系统间的协调性和灵活性,本文的 ETL 过程设计就是基于保险业通用数据参考模型而展开的。

3 基于数据参考模型的 ETL 过程设计和分析

数据仓库的一个普遍接受的定义是 20 世纪 80 年

① 该项目得到国家"十五"期间科技攻关项目(2001BA102A05-02)的资助

代由 Bill Inmon 提出的“面向主题的、集成的、随时间变化的、非易失的、用于进行战略型决策的数据集合”。数据仓库作为数据集成的中心点，是将数据转换为信息的第一步([2])，将数据提取到这个中心的过程就是 ETL——数据的抽取、转换和加载，ETL 过程涉及：访问数据；准备、清洗、校验数据；从各个系统中连结数据；转化数据以满足各种分析需求；加载数据到数据仓库中。

应用说明该模型在 OLTP 系统建设过程中发挥指导作用，可以加速软件数据架构的设计、清晰定义应用软件中各个业务组件间的关系和联系，同时，该通用数据模型也为保险行业的数据仓库设计提供指引，值得一提的是，当保险 OLTP 系统与 OLAP 系统采用相同的数据参考模型设计时，软件系统间的数据处理过程甚至 ETL 过程将变得非常的容易，本文论述的 ETL 过程就是以保险业通用数据参考模型为结构设计依据。

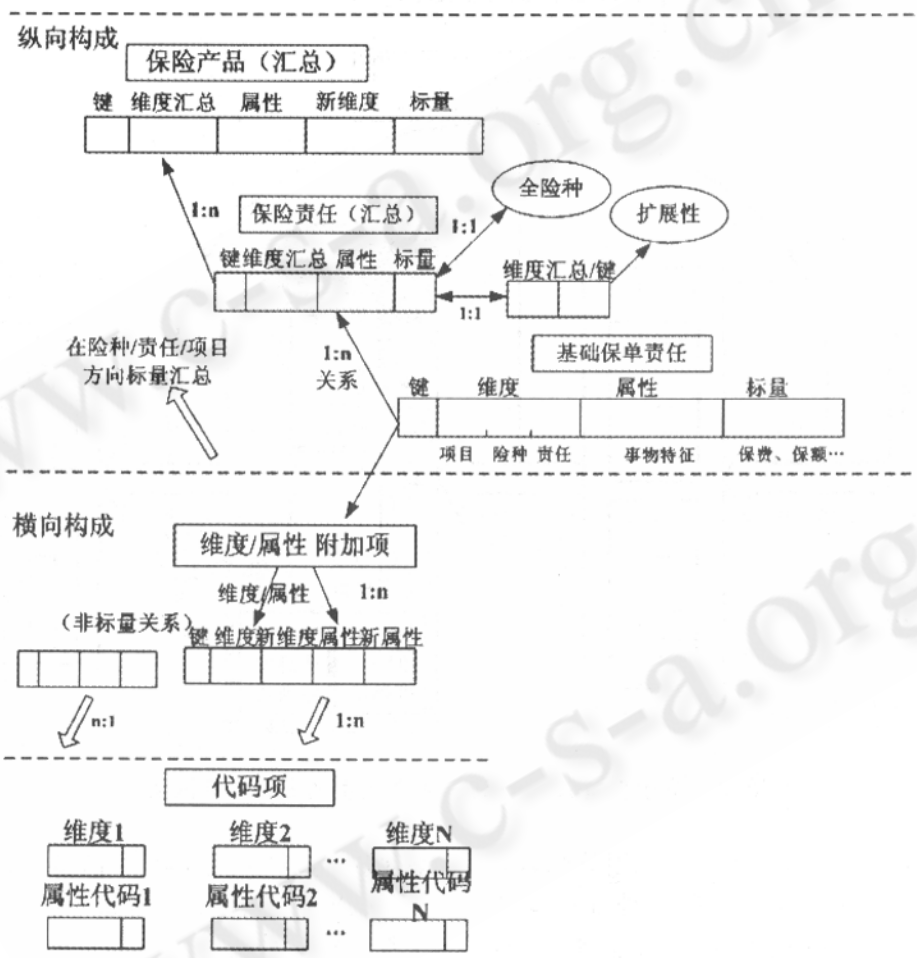

图 1　保险业数据参考模型简图

当前许多行业都设计开发了适用本行业/企业的数据仓库，而该数据仓库中的 ETL 过程却因为不能覆盖全部行业的特色而要么过于笼统、要么一笔带过；在实际的 ETL 过程中，规范化数据格式、代码转换、数据类型统一转换、数据整合、数据映射等部分经常困扰着技术人员。对于保险行业，如本文第二部分所述，技术工作者已经提出了通用的数据参考模型，并通过实际

本文设计的基于保险业通用数据参考模型的 ETL 应用架构如图 2 所示。

基于同一数据参考模型设计的操作型系统到分析型系统的 ETL 过程描述如下：

(1) 数据抽取(Extract)：从操作型源系统中去除操作型数据后，将必要的数据引入中间表。

(2) 数据转换(Transform)：根据分析型系统数据

结构的要求，插入、转换相应的时间元素、导出数据、相应的人工关系，如快照的有效日期，通过计算的总金额等；有的分析主题还需要改变粒度级别、合并相关数据表的操作；最终完成数据映射。

系统中的各个松散的组件联系起来，组成了一个有机的整体。而数据参考模型实例正是元数据管理中重要的数据定义依据。

在一般的 ETL 过程中，还都需要连接着日志总线。

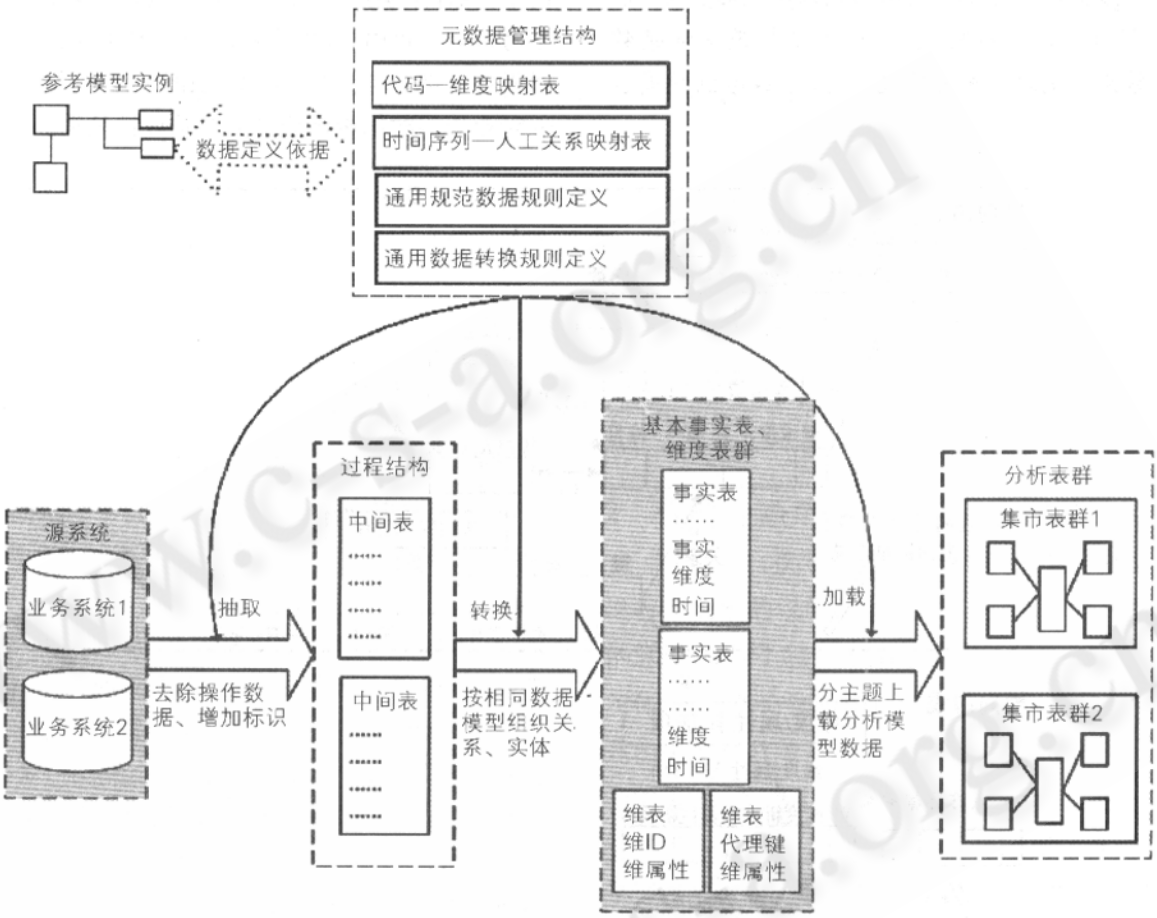

图 2　基于保险业通用数据参考模型的 ETL 应用框架

在此过程中，操作型系统纵向实体中的相关属性，加上分析系统中要求的时间元素、代理键后，转换为分析系统的事实表数据；操作型系统中的横向实体和代码定义作为分析型系统中的信息和维度定义，并通过增加相应的代理键后转换为分析系统的维度表数据。

(3) 装载(Load)：把表群中的数据按照一定的顺序装入到分析型系统的数据表中。

在以上的 ETL 过程中，元数据管理提供了数据资源的全面指南。元数据不仅定义了数据仓库中数据的模式、来源以及抽取和转换规则等，而且整个数据仓库系统的运行都是基于元数据的，是元数据把数据仓库

该总线中记录了错误日志、信息日志、检查日志等，为之后的核对、审计、回溯等工作提供依据。ETL 过程还关系到数据质量、效率等问题，因而在设计过程当中还需要考虑如下几个方面：

① 数据的准确性：数据源的歧义语义唯一定义和转化、缺失语义补充；

② 数据参照完整性：主动破坏后的重构、失败过程的回溯；

③ 过程的重用性、易调整性：随数据源增加、业务规则变化的适应性调整和兼容；

④ 提取的效率：合理的结构、优化的 SQL、高效的算法；

但我们可以看到，在基于统一通用数据参考模型设计的操作型系统和分析型系统的 ETL 过程中，以往花费时间最多的数据准确性、数据参照完整性，尤其是数据整合、数据转换规则定义的复杂程度大大降低。

对保单进行修正以适合最新的风险情况；在保险责任期间，不可避免地风险转变为损失，这样围绕损失保单关系人展开的一系列的交互工作，称为索赔－理赔过程。

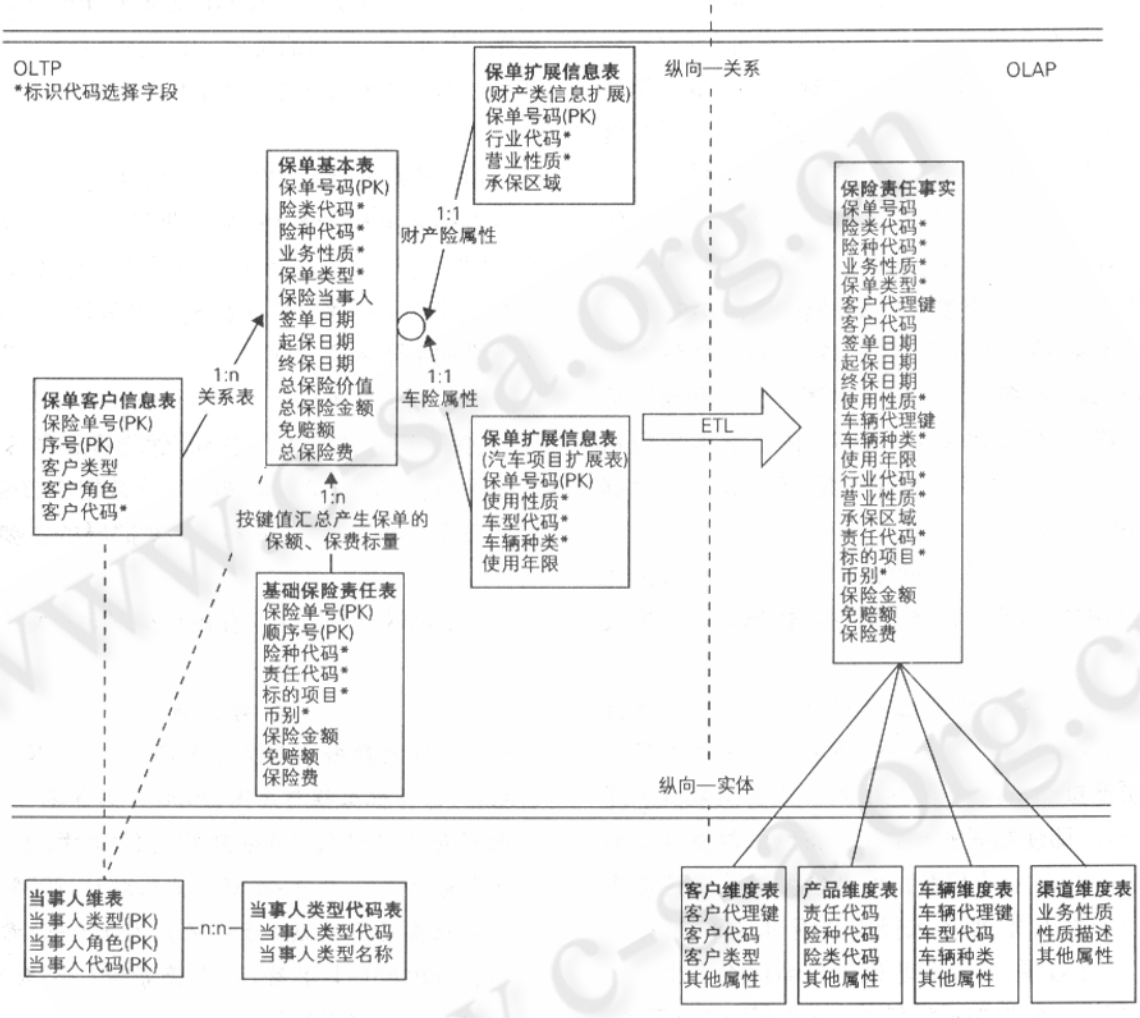

图 3　保险业通用数据参考模型指导下的两类数据实例及其 ETL 过程

4　保险公司的应用实例

ETL 处理的核心是数据，由数据描述的信息因各行业/企业性质的差异而各有各的特点和要求。保险企业是金融服务公司的一种特殊形式，它提供保户财产免遭损失的各种服务([4])，保险行业经营的产品不是一般的物品，而是风险，并且以契约的方式保证产品的售出和售后服务，这个契约就是保单；保单售出以后由于风险因素的变化，产生了保单所承担的风险范围、保险责任、责任免除等的变化，相应地，以批单的形式

在图 3 的结构中，我们采用保险业通用数据参考模型设计了操作型保险业务系统数据模型(以保单实务为例)，和分析型保单数据模型，并根据分析主题的需要使用 ETL 过程进行转化。在业务系统中，事务完整性是必须的、代码化处理则是提高处理性能的选择，以“纵向—关系、横向－实体”的分类方法可以清晰地看到纵向上的关系是保险经营的业务信息，而横向实体则是其中备选的代码(部分代码也并非用数据结构存储)。

在保单分析中，我们以保险责任为最小事实粒度，

将多对一的汇总关系映射到事实表中，同时将一对一的保单扩展信息中的关键属性加入到事实表中，在这个粒度最细的事实表上可以按某些/个维度进行汇总、衍生等处理等到其他主题分析；我们采用维度建模的方法将大部分的业务代码转换成平面维度表：保单客户信息表—>客户维度表、险类代码/险种代码/责任代码—>产品维度表、车型代码/车辆种类—>车辆维度表、业务性质—>渠道维度表，并根据设计需要对客户维度表和车辆维度表采用了代理键。

由以上描述可以看出，ETL 的转化过程将分为纵向、横向两类，横向处理关注于：① 业务通用代码转化为维度表、② 多对多层次化代码转化为平面维度、③ 代理键的替换；纵向处理关注于：① 一对一属性表关系的平面处理、② 多对一表关系的属性分拆和事实标量的运算、③ 代理键的引用。其中，由于参考模型相同，上述六个关注点的实现非常容易，而且保证了数据的质量和提取效率。

基于通用数据参考模型的 ETL 过程中，日常困扰着技术人员的数据不一致、代码转换、数据类型转换及数据映射过程大大简化。

在以往的实践中，保险公司的信息技术人员各自根据业务人员的需求编写程序，造成程序杂乱无章难以维护，更严重的是，由于程序员对于业务需求概念理解的差异导致即使相同的业务视角得到的数据结果不同；而在基于保险业通用数据参考模型的系统开发实践中，信息技术人员、业务人员等共同讨论需求、在原型上描述功能效果，并在此基础上遵循保险业通用数据参考模型设计分析型系统结构、开发 ETL 过程工具，在多系统的、异构的、分散的保险信息环境中建立起信息库，为保险公司掌握自身的发展情况、发掘市场机遇提供了有利的帮助。

5 结束语

数据抽取转化过程是 IT 中分行业/企业设计的技术，本文的设计技术所涉及到的数据建模遵循了保险业通用数据参考模型，由于保险业务系统与数据仓库系统采用了相同的数据参考模型，数据从操作型源系统转移到分析型系统的 ETL 设计过程和复杂度大大降低，本文详细划分了保险 ETL 的环节、ETL 分库设计的思路、ETL 回滚/回溯的想法。目前的设计方案中应该还可以有深入和改进的方面，也是下一步要做的工作，如：

(1) 元数据管理。在当前保险业的应用软件中，仍然存在很多异构的系统，多数据源定义、语义转换规则、缺失补充规则等在系统中的建立还比较浅显，如何建立保险行业中通用的元数据管理策略将是 ETL 过程完备性的主要考虑方向；

(2) 数据集市设计。保险行业模型涉及到人员和组织、保险产品、保险申请报价、保险费日程支付、索赔等诸多数据模型，如何构建和划分数据集市来满足保险行业多样性特点将是 ETL 深度扩展的主要考虑方向；

以上问题已列入下一步的研究探讨计划中。

参考文献

1 保险业通用数据参考模型及其应用，郑擎宇、郭妍、左春，计算机工程与应用，2006 年第 5 期.

2 数据仓库设计/(美)依默霍夫(Imhoff, L.) 等著；于戈等译，－北京：机械工业出版社，2004.12.

3 数据模型资源手册(修订版)，卷 1/(美)希尔瓦斯顿(Silverston, L.) 著，林友芳等译，－北京：机械工业出版社，2004.8.

4 数据模型资源手册(修订版)，卷 2/(美)希尔瓦斯顿(Silverston, L.) 著，林友芳等译，－北京：机械工业出版社，2004.8.

5 Building the Operational Data Store, John Wiley & Sons, Inc., New York, 1975.

6 http://www.ccidnet.com