保险业决策支持系统的数据仓库的设计与实现

Design and Implementation of Data Warehouse of Insurance Decision Support System

王文香 （中国科学技术大学信息科学技术学院 合肥 230027）
左 春 张 正 （中国科学院软件研究所 北京 100085）

摘要：本文设计了保险行业决策支持系统的数据仓库，详述了保险业数据仓库的体系结构、分析主题域、多维数据模型等，并对元数据库的设计、抽取转换加载过程的设计做了介绍。最后以在实践中取得良好效果的保险公司应用实例来说明数据仓库构建的 ETL 过程，信息展现的联机分析处理等。

关键词：保险 数据仓库 ETL 多维数据模型 联机分析处理

1 引言

本文针对保险行业的特点，设计开发了保险业决策支持系统的数据仓库，它完成从多数据源的原始操作数据中抽取数据，进行各种处理并转换成综合信息的过程，使加载到数据仓库中的数据具有良好的一致性、集成性。同时我们还提供功能强大的分析工具对这些信息进行多角度的分析，为保险企业进行全局范围的复杂数据分析、战略决策和长期趋势预测提供了有效的支持，从而能有效的进行风险防范，提供正确的运筹营销策略，提高市场竞争力。

2 决策支持系统数据仓库的设计

2.1 保险业数据仓库的体系结构

保险业决策支持系统的数据仓库的体系结构有四个部分：数据源、抽取转换加载工具、数据仓库、数据仓库信息展示工具，如图 1 所示。第一部分是不同的数据来源，包括业务系统数据库、收付费系统接口数据库、理赔数据库、再保险数据库、客户关系管理库等等；第二部分是数据抽取、转换、加载的过程，该过程完成从多个数据源中抽取数据，并对数据进行转换、规约，然后将整合好的数据加载到数据仓库；第三部分是企

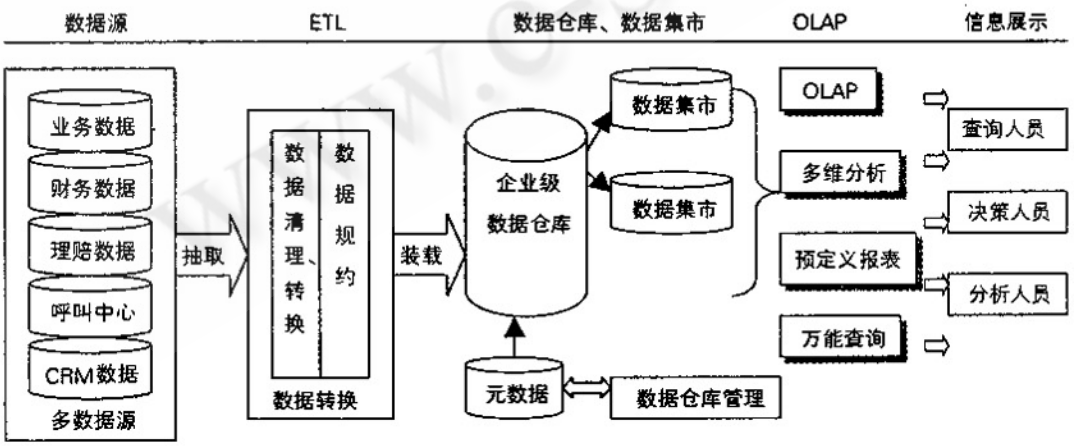

图 1 保险业数据仓库体系结构

业级数据仓库，为了更好的进行决策支持分析，数据仓库中的数据还可能被抽取到面向分析主题的数据集市；第四部分是联机分析处理，它能从不同角度、不同层次来观察分析数据，是给决策人员、分析人员进行信息展示的工具。本文在后面将详细说明形成数据仓库的 ETL 过程。

2.2 数据仓库的分析主题和数据模型

数据仓库中数据是面向主题进行组织的，主题是在较高层次上将企业信息源中的数据综合、归类并分析利用的抽象，每一主题基本对应一个宏观的分析领域。我们在清晰了数据仓库的体系结构之后就应该明确数据仓库的分析主题。按照保险行业决策支持系统分析的需要确定数据仓库的分析主题有：保单事件分析主题、客户分析主题、财务分析主题、理赔事件分析主题、营销分析主题等。

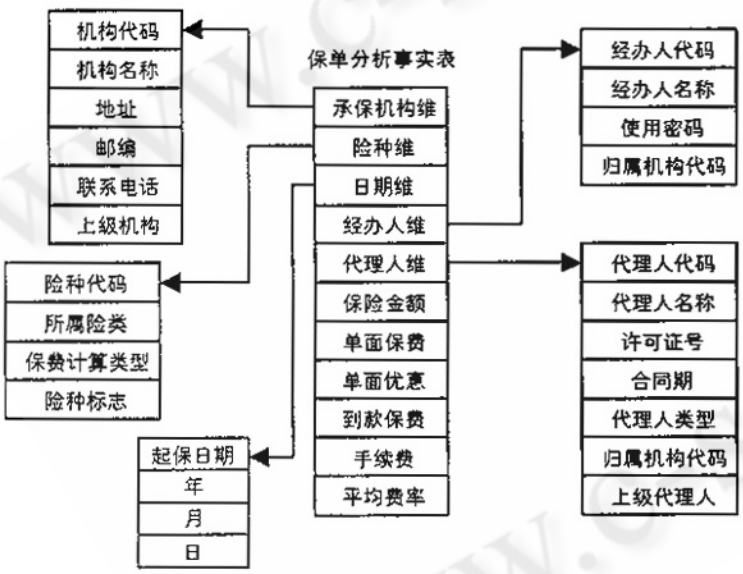

图 2　保单分析主题的星型模式

界定了数据仓库的决策类型，接下来考虑用何种数据模式来保存这些面向主题组织的信息。由于多维数据模型比较直观，易于理解，具有良好的扩展性能和快速查询能力[2]，已成为数据仓库普遍采用的数据建模方法。而星型模式是多维数据模型最常用的一种模式，它结构简单，建模方便，有助于优化数据仓库的性能，因此本文所论述的数据仓库采用了星型模式进行数据建模。星型模式由一个事实表和链接到该事实表的多个维度表组成。事实是决策者分析的目标数据，如保额、保费、手续费、赔款金额等。维是事实的信息属性，也是观察事实的角度，如保单的承保机构、险种、起保日期、经办人、代理人等。本文以保单分析事实表为例构造其星型模式如图 2 所示，我们可以汇总各个层次的承保公司的保费收入情况、手续费情况等，也可以从险种、起保日期、经办人等角度按照各种层次进行汇总。

清晰了数据仓库的分析主题和存储模式，下面说明管理数据仓库的元数据。

2.3 元数据库的设计

元数据是数据仓库的核心，它是描述数据的数据，用于存储数据模型和定义数据结构、转换规则、控制信息等。构建数据仓库的过程，即完成从各个系统中抽取数据，转化数据以满足各种分析需求，加载数据到数据仓库的过程，每一步都与元数据密切相关。

2.3.1 数据源元数据

数据源信息表，包括数据源的存储平台、数据格式、访问方法及使用限制、数据源的业务内容说明、数据源的更新频率、数据抽取需设置的参数、抽取的进度安排等信息。

2.3.2 预处理数据元数据

数据源映射表，确定从源数据到目标数据的对应规则，包括数据源的关系表和数据仓库主题表的复杂的多对多映射等信息；数据提取时间戳表，包括提数类型、提取的数据库及服务器名称、表名、本次提数时间、提数标志等信息；数据提取错误日志表，包括目标表名、错误类型、错误代码、错误信息、提数时间等信息；问题类型描述表；对比字段定义表，用于在通用的提数过程中，对比相同键值数据前次提数与本次提数记录是否发生变化的判断依据，包括表名、字段名、主键等信息；重复单号表，包括单号类型、业务号码、数据来源等信息；保单变更表，用来记录上次提取主题分析表到本次提取主题分析表期间发生变化的保单号，包括保单号码、分析表类型、提数时间、数据库及服务器名称等信息。

2.3.3 数据仓库主题数据元数据

数据仓库事实维关系表，反应数据仓库星型模式

的内部结构信息，包括主题名称、主题描述、事实属性、维表描述、维的关键字、数据层次及数据来源等信息；数据仓库索引、视图定义表等。

2.4 抽取转换加载过程的设计

数据仓库需要保证数据的正确性、一致性、完整性和可靠性（Correctness、Consistency、Completeness、Reliability）[3]。在构成数据仓库系统的诸多环节中，数据抽取（Extraction）、转换（Transformation）、装载（Loading）——通常简称为 ETL，是数据仓库系统中最基本且最重要的一部分，其性能的好坏直接影响到整个数据仓库系统的运行效率和最终分析结果。ETL 环节的处理过程如图 3 所示，在元数据的管理下数据从数据源到中间过程存储区、从中间过程存储区到中心数据仓、从中心数据仓到中心数据仓、从中心数据仓到数据集市，在各个处理的过程中都连接着日志总线，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。具体地说明如下：

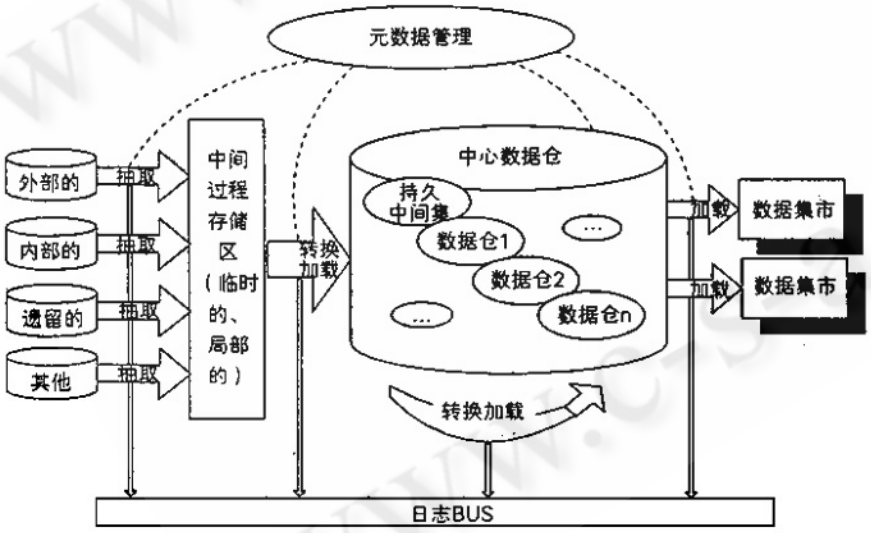

图 3　ETL 处理环节

数据抽取，是从数据源获取所需数据到中间存储区的过程。该过程会滤掉数据仓库中不需要的源数据字段或数据记录。鉴于保险行业的源数据量特别大，业务系统工作负荷重等因素，本文设计的 ETL 工具采用增量抽取的方式，即每次只抽取自上次数据抽取以来产生的增量数据和发生变化的数据。

数据转换，是对抽取到中间过程存储区的源数据根据数据仓库系统模型的要求，进行变换、关联、拆分或合并等处理。基本的转换规则有两个[4]：①一对一抽取，即源数据库中某一列直接抽取到数据仓库中的对应列，可分为直接抽取和通过聚集函数抽取两种情况。②多对一抽取，即两个或两个以上源数据库中的列被抽取到数据仓库中的一个列，通过算术运算或 Union 操作完成。本文设计的 ETL 工具除使用这两种基本的转换规则外，还编程自定义一些函数来满足复杂的转换需求。

3　保险公司的应用实例

3.1 ETL 过程的实现

保险行业经营的产品是以契约方式存在并承担风险的保单，在保险责任期间，风险可能会转变为实际的损失，这样保单关系人与保险公司就发生索赔－理赔的过程。保险产品和理赔本身所具有的生命周期特征为我们设计 ETL 逻辑分库和 ETL 周期提供了依据。因此在数据抽取转换加载的过程中，本文将数据在逻辑上划分为：已了保单库、未了保单库、已结理赔库、未结理赔库，用来区分保险责任或者理赔过程是否了结，这样便产生了历史的一般不变化的数据和活跃的数据之分。每一个 ETL 周期将增量产生数据，"未了保单库"和"未结理赔库"中的数据则采用增量结合刷新的方式产生，这样每次 ETL 所涉及的数量将有效地减少，从而缩短了 ETL 周期。为适应保险行业决策支持系统分析的实时性，ETL 工具每天晚上定时运行，完成从源数据库中抽取数据、转换、加载到数据仓库的整个过程。

以产生保单分析主题数据和理赔分析主题数据的 ETL 过程为例，来简要说明 保险行业数据仓库的 ETL 过程，如图 4 所示。其中：

DS1、DS2：DataSource，表示 ETL 过程的多个数据源；

MetaData：ETL 过程的元数据库；

PE：Preparing Environment，一个 ETL 周期中使用的基础数据准备环境；

PSE：Preparing Special Environment，一个 ETL 周期中使用的特殊数据准备环境；

RE：Running Environment，系统日常前端运行所依赖的基础数据库环境；

SE：Special Environment，系统日常前端运行所依赖的特殊数据库环境；

C＊_W，L＊_W：逻辑上的未了保单库，未结理赔库；

C＊_Y，L＊_Y：逻辑上的已了保单库，已结理赔库；

A：Analysis，增量提取的逻辑"主题分析库"；

synonym：同义名。

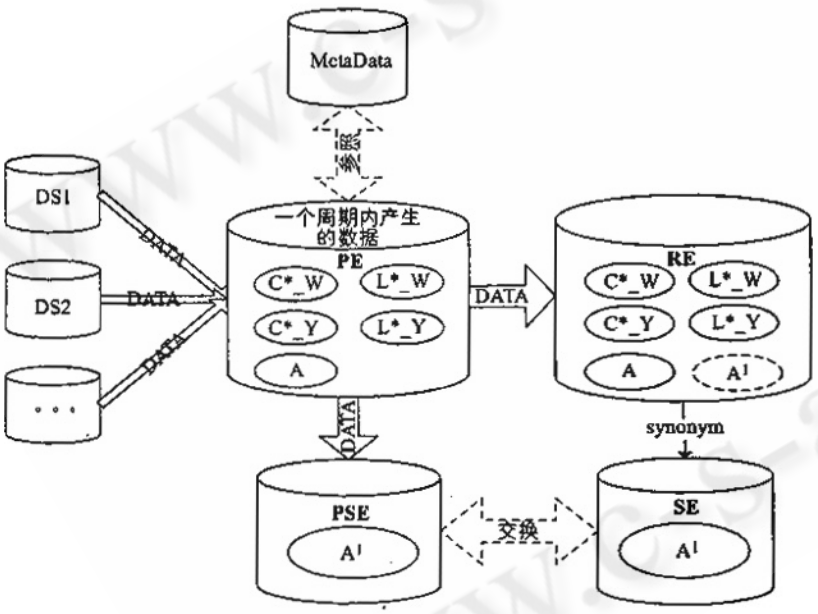

图 4 保险业数据仓库 ETL 数据处理过程

在每一个 ETL 周期，本文设计开发的 ETL 工具自动探测自上个周期结束之后保单业务数据、理赔数据的变化情况，ETL 将这些变化了的源数据抽取到 PE 中，然后 ETL 经过转换、加工处理产生目标数据集存储在 PSE 中，最后 ETL 将这个周期中转换的数据加载、刷新到 RE 和 SE 中。在以上的各个步骤中均有异常、错误等捕捉方法记录步骤中的异常，在每一步执行完毕后系统给出审计报告，由执行者决定是否继续执行，如果不继续执行系统进入回溯过程，依据异常记录重新处理数据。

3.2 联机分析处理

OLAP（OnLine Analytical Processing）作为一种多维分析工具，可提供数据多层面、多角度的逻辑视图。我们设计开发了基于 MVC 模式和 J2EE 多层架构的 B/S 模式的 WEB 展现方式，把数据仓库作为可用于分析的结果集，这些结果数据带有一定的粒度属性，这样具有汇总层次关系的数据集可以满足汇总分析的下钻和汇总角度的旋转，同时也可以满足对一定细节信息的查询。

4 结论

基于数据仓库的决策支持系统是一项综合性技术和解决方案，其目的是为企业进行全局范围的复杂数据分析、战略决策和长期趋势分析提供有效的支持。本文介绍的保险业决策支持系统的数据仓库已经在中国人民保险公司广州市分公司正式投入使用，结果显示，该系统很好地解决孤立多数据源、信息分散利用率低等问题。多维数据库模式的设计使得分析效率、分析能力大大提高，在一定程度上提高了公司的运作效率，并成为保险公司决策支持的有力助手。

参考文献

1 Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques[加]，范明、孟小峰 等译，机械工业出版社，2002。

2 武森、高学东，M. 巴斯蒂安，数据仓库与数据挖掘，冶金工业出版社，北京，2003。

3 Agosta L. The Essentia l Guide to Data Warehouse [M]. Prentice Hall，2000.

4 Anca Vaduva，K R Dittrich. Metadata Management for Data Warehousing：Between Vision and Reality [C]. 2001 Int'l Database Engineering & Amp；Applications Symp，2001.