农转资金项目综合数据仓库的设计与实现①

王庆芝 1,2 孙红敏 1 杨宝祝 1,2 张 俊 2 (1.东北农业大学 工程学院 黑龙江 哈尔滨 150030;

2.国家农业信息化工程技术研究中心 北京 100097)

摘 要: 农业科技成果转化资金系统是事物处理型系统，并不能对系统的数据进行快速有效的分析。为了有效地利用农业科技成果转化资金项目管理系统中所积累的大量数据，为农业科技成果转换资金的使用与效果进行分析，对今后资金合理分配做出决策，采用数据仓库技术实现了农转资金分析系统。系统设计并实现了从 E-R 模型向星型模型的转换，在此基础上，实现了农业科技成果转化资金项目信息的数据仓库。数据仓库以 B/S 模式进行数据展示，可有效实现农业科技成果转化资金项目信息的统一管理、统一展现，对随需而变的项目信息与项目绩效进行综合查询统计分析，以满足农业科技成果转化资金项目管理与绩效决策需求的不断发展变化。

关键词: 数据仓库；建模；ETL 过程

Design and Implementation of Integrated Project Data Warehouse of Agricultural Achievement Transformation Funds

WANG Qing-Zhi[1,2] , SUN Hong-Min[1], YANG Bao-Zhu[1,2], ZHANG Jun[2]

(1. Engineering college, Northeast Agricultural University, Haerbin,150030, China;

2. National Engineering Research Center of Information Technology in Agriculture, Beijing, 100097, China)

Abstract: The financial system of transformation for agricultural science and technology, a transaction processing-based system, cannot analyze data fast and effectively. In order to effectively use the large amounts of accumulated data, analyze the effect of the funds for the commercialization of agricultural science and technology achievements, and decide the rational allocation of funds in the future, an analytic system of agricultural achievement transformation funds is designed using data warehouse. System is designed and implements the conversion of the E-R Model to Star Schema, and the data warehouse of agriculture science and technology achievements transfer capital project was created on this basis. Based on B/S, the information centralization of management of agricultural scientific and technological achievements transfer capital project can be effectively achieved, and comprehensive query, statistical analysis of demand changeable project information, and project performance can be carried out to meet the needs of the constant development and change of agricultural scientific and technological achievements capital management project decision making performance.

Keywords: data warehouse; modeling; ETL process

1 引言

农业科技成果转化资金项目是从 2001 年开始的，此项目的事务型处理系统，已经积累了大量的数据，这些数据具有很大的利用价值，但是目前并没有充分的利用，要想充分的利用存在的数据，对农业科技成果转化资金的使用进行决策，就要对这些数据进

① 基金项目:国家科技支撑计划(2008BADB6B01, 2009BADA1B03)
收稿时间:2010-01-26;收到修改稿时间:2010-03-11

行重新的数据整合，建立数据或信息的共享平台。

数据仓库是近年来迅速发展起来的一种信息存储及管理技术，存储大量的、决策分析所必需的、历史的、分散的各种数据，经过处理将这些资料和数据转换成集中统一、随时可用的信息。本系统根据农业科技成果转化资金项目的实际需求，设计并实现农业科技成果转化资金数据仓库，可以满足灵活的查询和报表的生成，及时帮助政府了解农转资金所产生的社会效益和经济效益。通过数据分析，得出在每个阶段政策对农转资金使用的影响，辅助政府决策。

2 数据仓库总体设计

通过对农转资金数据仓库系统的设计，设计出系统的开发过程如图 1。

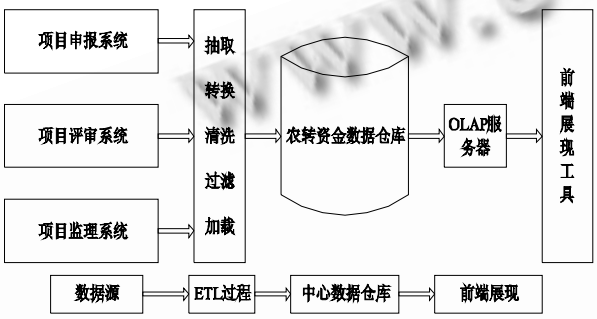

图 1　农转资金数据仓库数据流程

2.1 数据源

农业成果转化资金的业务系统分为申报子系统、评审子系统、监理验收子系统三个重要组成部分，每个部分各有一套数据库，数据具有一定的分散性和独立性，这为统计查询造成了一定的困难，因此需要将 3 个数据库进行整合，形成统一的数据仓库。

2.2 数据的 ETL 过程

数据仓库数据的获取需要经过抽取(extraction)、转换(transform)和装载(load)这 3 个过程,即 ETL 过程。这一过程是开发数据仓库的关键，占工作总量的 70%左右。在这个过程中，还要对选择的数据进行清洗和过滤，保证系统中数据的质量。

2.3 数据仓库和 OLAP 服务器

数据仓库不仅要存储通过 ETL 过程获得的数据，还要存储对系统的生命周期起着至关重要作用的元数据。OLAP 服务器按照多维数据形式进行存储，有利于查询和多维分析。

2.4 系统的模块

数据仓库的设计根据原有数据源和业务解决的问题来确定系统的范围和需求框架，通过对业务需求的深入了解，确定所要建立的系统包括项目数据整合模块、项目数据展示模块、项目数据报表模块、项目数据统计分析模块。

3 农转资金数据仓库的开发

数据仓库的构建是在业务系统上进行的，通过对系统的业务数据进行理解分析，组织数据仓库的主题，设计数据仓库的数据模型，实现数据的转换，并进一步完善数据仓库。

3.1 主题与数据模型的设计

3.1.1 主题的设计

数据仓库是对多个异构的数据源的有效集成，集成后按照主题进行重组。基于主题组织的数据被划分为各自独立的领域，每个领域有自己的逻辑内涵而又不相交叉。每个有效的主题域，都是根据业务需求而确立的。本系统需要对以下几个方面有决策需求：农转资金申报项目基本情况、农转资金申报单位基本情况、项目资金情况、经费来源、资金使用计划、农业科技成果转化资金项目执行情况、农业科技成果转化资金项目阶段指标完成情况、农业科技成果转化资金项目合同完成情况、项目执行期承担单位总体情况、项目执行末期科技成果达到的熟化程度、技术水平和市场前景。根据需求，确定相关主题分析：申报单位分析、申报项目分析、项目资金分析、经费来源与使用分析、项目社会效益分析、项目执行情况分析、项目完成情况分析、项目前景分析。

3.1.2 概念模型的设计

概念模型设计是在原有的业务数据库的基础上建立了一个较为稳固的概念模型。因为数据仓库是对原有数据库系统中的数据进行集成和重组而形成的数据集合，所以在设计数据仓库的概念模型时，首先要对原有数据库系统加以分析理解，然后考虑如何创建数据仓库的概念模型。概念模型既能正确反映用户的需求，又能反映现实世界，它最常用的方法就是实体-关系法(E-R 法)，其中用长方形表示实体，忽略 E-R 图设计中的关系。这里我们以项目社会效益分析为例来说明系统的创建过程。通过需求分析，画出业务数据的(E-R)图，如图 2。

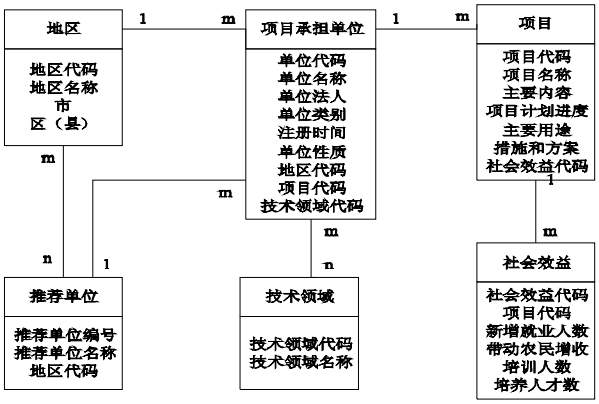

图 2　业务数据 E-R 图

3.1.3 逻辑模型的设计

逻辑模型的设计是数据仓库设计的重要阶段，即能反映业务的需求，又同时对系统的物理实施有着重要的作用。目前最流行的就是多维数据模型，它具有良好的扩展性和快速的查询能力，并且易于理解和多角度展示。多维数据模型是以星型模式、雪花模式和事实星座模式存在。星型模式是以事实表为中心，周围连接着维表，事实表含有大量的数据。雪花表示星型模式的变种，其中某些维表示规范化的，把数据进一步分解到附加表中，图形类似雪花形状。事实星座，多个事实表共享维表，这种模式可以看作是星型模式集。星型模式以增加存储空间为代价，提高了多维数据的查询速度。本系统采用星型模式，针对每个主题建立一个星型模型的多维数据集。要从以下几个方面入手。

(1) 确定度量、粒度。在多维数据值中，度量值是多维数据集中的中心值。比如，新增就业人数，培训人数等，它是用户最关心的数据。粒度的确定决定了数据单元的详细程度和级别，这里的粒度采用“最小粒度原则”，可以查询到每个项目、在每个时间段、每个地区、每个技术领域的社会效益。

(2) 确定事实、维度。以项目的社会效益为例进行分析，从业务数据的 E-R 图向逻辑模型转化，建立事实表和维度表，这里我们把项目所产生的效益作为事实，它包括各个维度的主键和一系列的社会效益，维度包括时间维度表、地区维度表、和项目维度表，其中项目维度包括项目承担单位、推荐单位和技术领域。各个维度都采用数值型的代理键，它唯一标识了每一个维度成员。更重要的是数值型具有连接效率高，便于聚合等优点。每个维度都是进入事实数据的入口，通过这种方法可以减少数据的遍历，提高查询速度。通过事实与维度的分析，社会效益的星型模型如图 3。

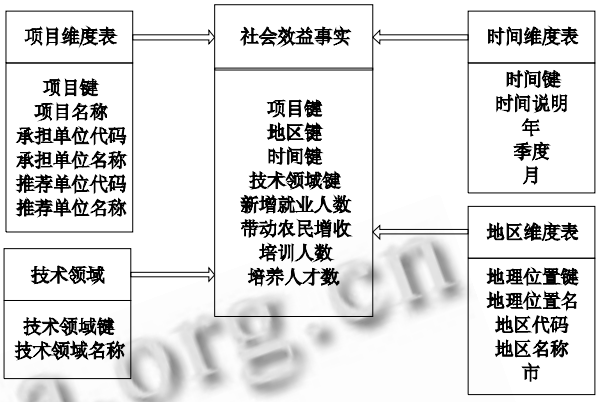

图 3　项目社会效益概况的星型模型

3.1.4 物理模型的设计

多维数据模型的存储主要有 3 种形式，MOLAP 模式，ROLAP 模式和 HOLAP 模式。ROLAP 利用细节数值存放于关系实际表格中，并将合计数值存放于关系数据库中；在 MOLAP 模式中，细节数值和合计数值均存放于立方体中；而在 HOLAP 中，将细节数值存放于关系实际表格中，而将合计数值存放于关系数据库中。三种模式各有各的优势，在本系统中选用 ROLAP 模式，这种模式易于管理，能够保证数据安全性和完整性。

3.2 系统模型的实现

农转资金数据仓库系统的服务器采用 Windows 2000 Server，数据库服务器采用 SQL SERVER2000。此系统采用 B/S 结构进行数据展现，系统用 J2EE 架构实现。WEB 服务采用 Tomacat(可以移植)。

数据的 ETL 过程一共分为抽取、转换、加载 3 个过程。先把数据抽取到准备区域，然后根据数据模型有目的的选择源数据，转换过程中需要对数据进行清理、过滤、集成。之后再加载到数据仓库中。在这个过程中，制定相应的抽取规则，按照抽取规则和数据模型进行数据从源数据到目的数据的转换。抽取过程使用 SQL SERVER 2000 中的 DTS 作为 ETL 的开发工具。在进行数据统计分析时，使用功能更加强大的 MDX,可以对数据进行多角度钻取、切片、旋转等分析。下面选取了社会效益指标的年度、地区分布和技术领域三个维度进行钻取。通过点击如图 4 的图表中的加号和减号，可以进行灵活上钻下钻。

图 4　社会效益钻取报表

4　总结

本文在农业成果转化资金项目管理系统的基础上，建立了相应的数据仓库。数据仓库的构建，是一个不断循环的的过程，通过不断的完善与维护，最终形成完善的数本文在农业成果转化资金项目管理系统的基础上，建立了相应的数据仓库。数据仓库的构建，是一个不断循环据仓库。通过对业务的分析，构建了了符合农业成果转资金分析的数据仓库，为农业资金的使用及效果作了全面而系统的分析，对全国的农业给予最优的支持，成为农转资金在农业各个领域的合理分配有利工具。

参考文献

1 饶元,冯博琴.数据仓库的存储优化设计.计算机工程与应用, 2003,36:198.

2 李静,陆菊康,汤艳艳,章铭. 基于数据仓库的销售分析系统的设计与实现, 2003,29:174.

3 于宁,王行言,罗念龙.高校教学决策支持系统数据仓库的研究与实现, 2006.10.

4 陈伟民.数据仓库与数据挖掘教程.北京, 2006,20:3854-3855.

5 Chen MC, McNamee LP. The data model and access method of summary data management. IEEE Transactions on Knowledge and Data Engineering. 1989.

6 Inmon WH. Building the Data Warehouse. New York: John Wiley & Scons Inc , 1992.

7 Kelly S, Sach T. Data Warehouse Administration, SAS Institute White Paper, 1997.

8 Agosta L. The essential guide to data warehousing. Prentice Hall, 2000.

9 Shapiro JR. SOL Server2000 参考大全.北京:清华大学出版社.