关于建立税务数据仓库的思考

龚振中　黄　敏

（中南财经政法大学财政税务学院　武汉　430060　湖北鄂州市国家税务局　湖北鄂州　436000）

【摘要】本文分析了我国税务信息化建设中存在的问题，并在此基础上提出建立税务数据仓库，同时构建了我国税务数据仓库的体系架构，以提供借鉴。

【关键词】税务信息化　数据仓库　数据挖掘

数据仓库于上世纪90年代开始流行，经过十几年的发展，数据仓库技术已成为企业和公共管理部门信息集成、信息资源开发和决策支持的最佳解决方案，但数据仓库技术在我国的应用还处于起步阶段。2006年我国国家税务总局正式实施以“覆盖所有税种、所有工作环节、国地税局并与有关部门联网”为目标的金税“三期”项目，笔者认为应以此为契机，在一体化原则下，在全国范围内有计划地建立和推行税务数据仓库项目，提高税务征收质量和效率。

一、应当建立税务数据仓库

1. 税务信息化建设存在的问题。税务部门信息化建设经过多年的发展，在减少税收执法随意性、加强税收管理和监控、打击偷逃骗税等方面取得了显著成效。但是，从一体化建设的总体要求来看，还存在以下问题：

（1）应用系统分散。由于在信息化建设初期缺乏统一规划，税务部门开发并使用了许多基于不同的工作平台、代码各异、各自独立的应用系统，各应用系统之间的信息无法全面衔接、数据缺乏共享、功能存在重叠，从而增加了征纳双方的工作量和税收成本。

（2）信息质量不高。税务部门的办税窗口服务实行的是登记制而不是审核制，税务部门无法核实纳税人的申报资料，从而必须到企业实地核对后才能发现问题，加之税务人员在数据采集、修改、传输过程中又会发生错误，从而将会产生大量的垃圾数据。

（3）信息应用效率较低。在日常处理数据库中，存储着大量税收业务的基础明细数据，这些业务数据隐含着十分丰富的信息和规律，但是难以直接供管理者和决策者使用；不同系统的查询模块的口径不一致，大大降低了税务信息的应用效率。从总体上看，税务部门的数据应用仍停留在较低的层次上，如简单的信息查询和基础数据的比对，无法实现多维复杂信息查询和知识发现，不能满足纳税评估、稽查选案和税收决策的需要。

2. 建立税务数据仓库的重要意义。

（1）全面整合信息资源。数据仓库通过一体化规划，全面整合税务信息，通过跨系统的取数来实现数据的高度共享，将各类涉税数据整合到系统中，从而消除税务部门的“信息孤岛”现象。数据在进入数据仓库前，经过抽取、净化和转换，其正确性、一致性和完整性得到了加强。通过对业务系统的数据进行综合比较、测评和分析，用户可以发现日常操作和传输过程中产生的错误数据，从而及时进行修正，消除信息传递过程中的信息失真和信息弱化。

（2）加强信息的可用性。数据的集中存放使用户查询数据只需要访问中心数据库，无须在不同的数据库系统之间进行切换，使得用户存取和使用数据更为方便。数据仓库数据的呈现方式丰富、灵活，包括图表、地图等，能满足各个层级用户的使用偏好。基于数据仓库的数据挖掘技术和联机分析处理系统（OLAP）从海量数据中提取隐含的有用信息，帮助税务部门对数据进行微观、中观和宏观的统计分析，使其科学而高效地进行决策。

二、税务数据仓库体系的基本构想

（一）省级数据仓库体系构建

省级数据仓库兼具中观、微观两个层面的应用，立足于税务部门的综合数据分析需求，坚持一体化原则，实现“辅助决策、服务征管、全面监督、优质服务”的目标。

1. 主题的选定。主题是在较高层次上将信息系统中的数据综合、归类从而得到的抽象概念，每一个主题都是决策者所关心的问题，具体来说，主题对应的是部门中某一宏观分析领域所涉及的分析对象。

省级税务分析对象的主题可以设计如下：第一类是“税收宏观变化分析预测”，可进行宏观经济分析、税收分析等税收和国民经济的综合评价工作，挖掘税收经济的潜在运行规律，研究税收制度变化对相关经济指标的影响，为税收制度改革提供第一手资料；第二类是“税收报表、查询分析”，用于定义或编制本部门所需的日常的和特别的统计报表，对税务部门以及本部门的各项工作指标进行考核；第三类是“稽查、征管、发票分析”，为诸如稽查选案、征管业务分析提供支持，根据用户交互系统的需求，经过分析或挖掘将纳税人异常税务行为、纳税人流失、专项整顿目标纳税人等信息反馈到各操作型系统，发现日常税收业务活动中的漏洞，产生预警信息，供纳税

服务大厅工作人员和专管员对相应纳税人做出更具针对性的决策判断;第四类是“纳税评估、纳税人知识分析”,纳税评估不仅包含通常所指的纳税评估,还包含对税务登记、税务稽查等单位涉税信息的综合利用,而纳税人知识分析包括纳税人价值分析、纳税人行为识别、纳税人信誉度评级等内容。

2. 维度及分析指标的确定。维度是分类的、有组织的层次结构,用以描述数据仓库实事数据表中的数据,在多维数据集中对度量值进行分类,以便于分析,每个维度表的主键都与多维数据集的实事数据表或另一个维度表中的外键联接。税务数据仓库的维度从总体上可以分为时间维、区域维、税种维、税务机构维、纳税人维、专项业务维、日常业务维七种类型。具体维度根据分析主题的不同而有所不同,以发票分析多维数据库为例,建立的维度对象主要有日期、行业类别、地域类别、票种、税务机关、纳税人等。数据分析指标是业务分析和监控使用的指标或量度,操作人员用分析指标来衡量表现情况。税务数据仓库的分析指标可以分为基本指标和衍生指标两大类。基本指标包括税额、纳税人数量、发票发售份数等绝对指标;衍生指标包括平均指标、相对指标、比例指标、结构指标、比较指标,如纳税人实际税负率、计划完成进度等。

3. 数据仓库的建设。在规划数据仓库体系架构时,需要为体系的每一个组成部分设计工作模式,将规划所有的组成部分作为一个整体工作,税务数据仓库体系架构包括税务数据获取、税务数据存储、税务信息传递三个主要区域。

(1)税务数据获取。包括涉税数据源的选择、数据源信息的抽取、将抽取出的数据移入数据准备区域,以及向数据仓库载入数据作准备的整个过程。数据源包括税务部门的内部业务数据(综合征管系统与出口退税系统等系统数据)、外部数据(包括工商与地税等单位涉税信息)和历史数据(以前曾使用系统所采集的业务数据)。

数据抽取、转换、清洗和载入改造了源系统中的相关数据,将它们变成有用的信息并存储在数据仓库中。从源系统中抽取的数据主要有静态数据和修正数据两类。静态数据一般在数据仓库的初始装载时获取,修正数据的抽取可能是立刻进行或者延缓进行。数据的转换和清洗就是通过对来自于源系统中的数据进行格式修正、度量单位的转化、键的重新构建等来提高数据的质量,包括对已抽取数据中的缺失值进行补充。在数据准备区中完成数据的转换和整合后,就要将准备区数据向数据仓库储存库载入,数据装载分为初始数据加载和日常数据加载。初始数据加载是在完成设计和建设数据仓库的工作后的初次数据加载;日常数据加载是在操作数据变动后,按照一定规划将数据存入正在工作的数据仓库中。

(2)税务数据存储。税务数据存储环节是整个数据仓库环境的核心,其存储分析所需的大量历史数据,提供对数据检索的支持,具有支持海量数据存储并能快速检索的突出优点。数据仓库数据组织的构建可采用自上而下的方法,即先建立一个企业级数据仓库,然后将其中的数据加载到各部门的数据集市和不同主题的数据集市中去。

(3)税务信息传递。这个过程设计很多不同的向用户传递信息的方法,使得用户可以从企业级数据仓库或独立的数据集市中更便捷地访问信息,数据流从企业级数据仓库或独立的数据集市流向各类应用系统。数据仓库提供OLAP支持,OLAP提供了概括化和信息下钻的功能,既能概括化到较高层次的聚集,也能下钻到较低层次的细节。多维数据库从基本数据仓库载入数据,并以多维信息立方体的方式保存,用户通过多维数据库中的这些信息立方体来执行复杂的多维查询,通过对立方体的旋转,用户能够看到立方体中各种不同切片的页面显示,多层次、多视角的数据查询功能帮助用户更好地理解这些数据并做出正确的判断和决策。

(二)各级数据仓库之间的关系

依照现行机构设置和职责,总局数据仓库立足宏观视角的决策和分析预测,省级数据仓库立足于中观和微观视角的政策层面和业务决策分析。各省级数据仓库之间不互相联通,各省在各自范围内查询本省的数据资源,需要跨省的数据分析由总局数据仓库完成。总局数据仓库通过从基层操作型系统直接取数的方式,转变为从各省级数据仓库中抽取所需的各种数据,这样一方面能提高总局数据仓库的数据处理能力,另一方面也可以降低网络带宽的压力。

(三)操作型数据库与数据仓库之间的关系

操作型数据库中存放着大量涉税基础数据,是数据仓库系统的数据来源,数据仓库经过抽取、清洗等环节,按主题存储,形成部门内部有关涉税资料的统一数据平台,并根据客户交互系统的需求,经过分析或挖掘,将分析结果反馈到各操作型数据库,产生预警信息,辅助用户进行分析和决策。

三、建立税务数据仓库应注意的问题

1. 严格执行项目管理。因数据仓库涉及的部门多、项目范围广、问题复杂、技术全新,只有严格执行项目管理、充分考虑项目开发的特性,才能保证项目的成功上线。

2. 确保数据质量。建立数据质量责任制,明确质量小组人员在数据清洗过程中的角色和义务,落实数据清理责任;确定数据质量的审核标准,符合标准的数据才能进入数据仓库,保证基础数据的准确性。

3. 业务需求驱动数据。使税收业务需求而不是技术成为数据仓库的驱动力量,关注用户需要什么样的信息,通过了解用户的总体需求,建立支持用户需求的、统一规范的税务数据指标体系。

4. 建立严密的安全体系。总局和省局两级数据仓库的建立、运行还需要备份和运行维护系统的支持,应相应配备两级灾难备份中心和两级维护支持中心,保证系统运行与维护管理工作的顺利开展,确保各级数据仓库的信息安全。

主要参考文献

1. 夏火松.数据仓库与数据挖掘技术.北京:科学出版社,2004

2. 国家税务总局信息中心.推行“一户式”管理 深化信息资源整合.中国税务,2004;11

3. 周根贵.数据仓库与数据挖掘.杭州:浙江大学出版社,2004