学校编码：10384　　分类号_____密级_____
学号：X2008230230　　UDC_____

厦门大学

硕　士　学　位　论　文

基于数据仓库的新疆国税税收分析系统的设计与实现

Design and Implementation of XinJiang Tax-data-analysis System Based on DW

李　扬

指导教师姓名：董槐林教授
专　业　名　称：软件工程

论文提交日期：2010 年 10 月
论文答辩时间：2010 年 11 月
学位授予日期：　　年　　月

答辩委员会主席：_____
评　　阅　　人：_____

2010 年 10 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外,该学位论文为(　　)课题(组)的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（ √ ）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

随着税收征管体制改革的不断深入和税务部门信息化水平的不断提升，以信息化为支撑的专业化税收征管格局已经形成。新疆国税信息化建设经过十几年的不断深入发展,随着税收数据的省级集中,税务部门积累了大量富有价值的电子数据,但是大部分数据处于粗粒度、零散、集中整合度低等不利状态，很难为管理层的决策分析提供有效的支持。为了更加科学的利用这些数据资源,充分挖掘税收数据的价值，建立基于数据仓库技术的税收分析系统势在必行。

论文立足于新疆国税的实际情况，以数据仓库技术为基础，结合联机分析处理和数据挖掘技术，采用 B/S（Browser/Server）模式，设计并实现了新疆国税税收分析系统。系统利用 ETL 工具对综合征管软件、防伪税控、出口退税等系统中的税收数据进行抽取、清洗、加工，形成了基于统一技术框架的基础数据源，完成了新疆国税税收数据仓库的构建以及 OLAP 应用的构建。

通过构建新疆国税税收数据仓库以及联机分析处理技术的应用，建立的新疆国税税收分析系统，能够帮助新疆国税各级管理者提高各项政策制定的科学和理性，为税收领域的税源监控、科学化管理、数据分析、风险分析以及税收预测提供有效手段。目前，以本文设计并开发的新疆国税收分析系统已在新疆国税的各级税务机关中测试使用，收到了较好的效果。对各级管理者准确把握税源动态，实现对整个税收活动的有效管理和监控，具有重要的意义，为税收分析决策提供有利保障。

关键词：数据仓库；联机分析；税收分析

Abstract

With the reformation of taxation system and the upgrade of information technology used in taxation bureau, a specialized taxation management pattern which is based on information technology has been established. Large numbers of valuable electronic data has been accumulated in tax department during the past ten years. But most of the electronic data can not be very supportive for decision-making for its rough, random and disorder. So, it is necessary to set up a tax-data-analysis system which is based on a data warehouse technology to make the most use of these date resources.

According to the reality of Xinjiang national tax bureau, the author design a tax-data analysis system in Browser/Server mode by using data-storage technology, online analysis and data excavate technology. The system can get data and process data from CTAIS and export drawback software by using ETL. By the system, we can build Xinjiang national tax database and realize OLAP application.

The system will do great help to decision-maker to make more scientific and rational policy which is effective method to enterprise monitor, scientific management, data analysis and tax predict. At the present time, the system has been testing used in all levels of taxation department in Xinjiang. It is of important significance for governor to monitor enterprise and make right tax analysis decision.

Key Words: Data Warehouse; Online Analytical Processing; Tax-data Analysis

目 录

Contents

第一章 引言

1994 年我国进行工商税制重大改革以后，税收信息化建设得到了高速的发展。从 1994 年初开始实施“金税工程”一期试点工作以来，到 2000 年的金税二期工程的完成，以及金税三期的顺利实施建设，税务机关特别是国税机关的税收信息化建设已初具规模，随着税收数据的省级集中，各级各类应用系统虽然积累了大量的基础数据,但是大部分数据都处于粗粒度、零散、集中整合度低等不利状态，无法快速转换为决策信息。随着税收信息化建设的进一步深入和数据仓库技术的迅猛发展，为开展深层次的数据分析、提升数据应用效果提供了可能。本章主要介绍了基于数据仓库的新疆国税税收分析系统建设的背景和意义，从税务信息化建设发展的现状以及存在的问题进行了阐述，同时也对本文研究的内容以及本文的结构安排等进行了总体概述。

1.1 研究背景及意义

随着税收信息化建设的不断深入，以综合征管信息系统在全国各省、市、自治区国税系统推广应用为标志，我国税收管理信息化发展已由初始化阶段、传播阶段、控制阶段，成功地进入集成阶段。形成了以增值税专用发票稽核、协查系统、防伪税控系统、综合征管软件、出口退税系统、多元化电子申报纳税系统和办公自动化等为主要应用的信息化格局。税收数据的省级集中有力地促进了税收业务的整合和工作流程的优化，减少了数据传递的中间环节，增强了上下级数据信息的对称性，提高了对税收质量和执法行为的监控，给数据的分析利用积累了大量的基础数据，为开展深层次的数据分析、提升数据应用效果提供了可能。数据分析利用逐渐成为税收信息化工作的重点，建设税务数据仓库成了税收信息化建设的发展趋势。国家税务总局在金税三期规划中，也将建设数据仓库和决策支持系统列为重要内容。新疆国税税收分析系统将各个孤立应用系统中的异构数据经过过滤、清洗、抽取加工和重新规划，形成了统一技术架构，数据覆盖税收征管、防伪税控、稽核、进出口等业务系统的基础数据源，充分利用海量的业务数据，进行深入挖掘，以分析、预警、查询统计、报表等为应用主导，运用图表结合的丰富展现形式，使信息系统真正成为管理

决策的有力工具。税收分析系统能够使原本分散的税收基础数据有效地实现综合分析和增值利用，为信息资源管理和税收分析决策提供有利支持。其重要意义如下：

(1) 有利于把握收入进度和收入趋势，更好的完成税收任务。通过加大分析力度，税务部门可以准确掌握企业生产经营形势和资金运转情况，可以有效地把握影响全局收入的重点环节；

(2) 可以通过测算税负水平来评价征管水平。通过对同行业纳税人的税负进行测算，得出该行业的平均税负水平，再根据该行业中个体税负与行业平均税负水平的离散程度，可以有效的反应出各地征管的差异水平；

(3) 有利于增强税收计划编制的科学性。通过加强税源分析，可以比较透彻的摸清底数，在此基础上编制的税收计划就会比较客观、真实，比较符合各地实际，有利于增强收入计划编制的科学性，促进依法治税；

(4) 可以为加强税收征管提供数据支持。通过税收分析，我们可以从掌握的同业税负在不同地区的差异、入库率、税收弹性或征收率等多种量化指标中发现征管工作存在的漏洞与薄弱环节，为提高征管质量指明方向；

(5) 可以为纳税评估、收入预警提供参考标准。通过对企业有关财务数据和生产经营情况的掌握和了解，可以分析出企业税源变化规律；

(6) 能够通过贴近式管理，寓服务于管理之中，结合纳税人的实际需要提供优质的服务，有效地解决“疏于管理，淡化责任”的问题，真正将信息资源管理和优化服务有机的结合起来。

1.2 国内外研究动态

在实际运用领域中，传统数据库的联机事务处理(On-Line Transaction Processing，OLTP)经过较长时间的发展已达到一个基本成熟的阶段，而联机事务分析处理(Online Analytical Processing，OLAP)、数据仓库（Data Warehouse，DW）和数据决策系统（Decision Support System，DSS）则正处于成长阶段，并在逐步迈向成熟[1]。

在一些欧美国家，以数据仓库为基础的联机分析处理和数据挖掘应用，已经在金融、保险、证券、电信等传统数据密集型行业取得成功，在税务行业的

典型的案例也有很多，如：

IBM 公司帮助新西兰国税实施了 CRM；1998 年帮助加州税务启动了基于 IBM DB2 数据库软件的综合逃税人监察项目数据仓库解决方案（INC）项目，使加州税务能够在超过 2.2 亿项的独立税务信息中利用商业智能技术进行业务分析。

NCRTeradata 已经成功地实施了包括美国国家税务局（IRS）、澳洲国家税务局（ATO）等在内的数据仓库和数据挖掘项目。数据仓库的效益仅 1996 年就帮助美国国家税务局追回补交税款两亿笔、增收 200 亿美元的税金和罚款，并进行了 120 万笔帐目审计。CRIS 系统已成为美国国家税务局当前和未来实现税务目标的重点。

对于数据仓库技术的运用，我国远迟于欧美国家。但随着该技术的发展，在国内各行业中的应用也逐年提高，一些优秀的数据仓库技术也逐渐进入到税务领域中。许多大公司针对税务系统均有各自解的解决方案问世，但大多借鉴或直接使用其较为成熟、完整的国外产品。

湖北省地方税务局的数据仓库系统选择了 Sybase 的数据仓库解决方案。项目前期于 2003 年 12 月由 Sybase 数据仓库服务部来负责具体实施，二期则交由 Sybase 的合作伙伴 Bestinfo 公司数据仓库事业部承建。通过两个阶段项目的建设，目前已建成了数据仓库中有关纳税户、税金、税源普查和社保普查四个业务主题。并在数据仓库系统的基础上，实现了面向主题的联机分析系统，包括纳税户等四个主题的即席查询、数据钻取、多维分析。

国内的税务数据仓库建设项目，还有 IBM 公司为天津地税、武汉地税、西安地税、北京地税等提供了不同的解决方案；青岛国税的征管系统则采用了 Oracle 产品，目前已经完成；浪潮通过与中国人民大学金融与财税电子化研究所长期合作的浪潮纳税评估系统，凭借在税务行业信息化领域积累的丰富经验，基于纳税评估的平台建设、数据采集、模型和指标的建立、工作流等方面推出的一套完整的解决方案。

以上是国内外税务分析支持系统中部分较为前端的解决方案，各有其优点，在技术上也较成熟。但是也存在诸多问题，主要表现在如下几个方面：

（1） 系统适应能力不强

各地在构建数据仓库产品时往往没有很细致地研究税收业务的特点，比较盲目地引入国外厂商提供的集成数据仓库产品以及数据模型。由于国内税收业务比较复杂多变，新的业务调整需求层出不穷，实现新的需求涉及数据模型、ETL 、前端应用等一系列调整，容易使系统陷入疲于应付的状态，对业务的全面支持尚待发展。

（2） 信息表现方式单调

目前，数据仓库应用主要以固定报表、OLAP 分析为主。固定报表一般是将原来由手工或者半自动处理的管理报表在数据仓库的基础上再重复一次。OLAP 分析是从不同的角度观察指标，并不是尽善尽美的信息表现手段。这样既不能完全支持分析的要求，也远远没有发挥出数据仓库的应用价值。

（3） 系统不够稳定

由于税收信息化建设发展迅速，作为主要数据源的核心业务系统始终处于升级、换代的过程中，有些甚至是颠覆性的变化（如从 SYBASE 数据库迁移到 ORACLE 数据库，从英文字符集变更为中文字符集），造成原有设计必须进行重大调整，甚至重新设计。

（4） 系统总体成效不高

由于开发缺少统一的规划，各地重复开发，系统各具特点，技术结构、业务统计口径不一致，本地实现的功能无法到其它地区应用，不具备全国推行的条件。由于数据仓库的建设投入的成本都比较高，造成总体成效不高。

1.3 主要研究内容与目的

本文研究的主要内容是从新疆国税实际出发，按照整合现有数据资源，提高税务数据的应用率，提高税源管理水平，促进税务决策科学化的指导思想，应用数据仓库技术，将税务工作中生产的各类业务数据进行归集、清理、转化，建立一个架构灵活、易于拓展、操作简便、功能强大的税收分析系统，从而实现从多角度对税收数据进行多维查询和分析，通过这些查询和分析的结果，税务部门的业务人员可以很直观的分析出个企业、地区、行业的税收特点，据此准确地为各级税务部门提供决策依据，有效的指导税收工作。系统的主要功能模块包括查询统计、分析监控、报表管理等。

本文研究的目的在于：

（1） 通过建立数据分析指标体系，对日常征管中进行数据分析利用的指标进行整理、分类，逐步建立一个包括税负分析、税收管理质量分析、发票管理分析、税收执法责任考核、关联企业监控分析、数据质量监控分析等方面的数据分析指标体系，解决目前数据分析中常见的“分析报告成果不共享”、“分析指标口径不统一”、“已解决问题死灰复燃”等问题。；

（2） 通过构建数据处理分析模型，充分利用先进的数据库技术、多功能的数据展现软件、数据挖掘工具创新分析方法。引入经济学、统计学的有关数学模型，参考国内外在数据分析方面的做法，提高对数据的挖掘和综合分析能力。对各个应用系统后台数据库中的相关数据进行抽取、过滤、关联、整理和比对，构建符合征管业务特点的标准化数据处理分析模型。这些模型综合运用了关联分析法、趋势分析法、定量和定性分析法、归纳推断法等分析方法，为数据处理分析工作提供了有效的抓手。；

（3） 通过搭建数据仓库，根据数据分析主题和主要业务数据主题，结合业务需求，通过抽取工具从征管以及相关系统数据库中，按照主题抽取建立“关系数据存储（ROLAP）”，再建立通用数据分析指标和分析模型，组织基于关系数据存储（ROLAP）的统计、分析类应用功能设计、开发、试点和推广。与此同时，搭建数据仓库平台，统一定义元数据标准，建立和完善“多维数据存储（MOLAP）”平台，建立基于数据仓库的知识库、决策分析模型；基于数据仓库和数据集市进行数据挖掘，实现分析、决策类的复杂税收分析应用功能，并及时提交给相应的管理部门，供管理部门作出提高征管效能的决策。

1.4 本文的组织结构

本论文共分为五章：

第一章 引言，对论文的研究背景、意义、国内外研究现状以及主要研究的内容和目的进行了详实的分析和评述。

第二章 相关技术，介绍了新疆国税税收分析系统开发的主要运用的技术，如：数据仓库、数据集市、联机分析处理等。

第三章 新疆国税税收分析系统的需求分析，主要描述系统应用的源数据、