学校编码：10384　　分类号_____密级_____
学号：X2009230317　　UDC ____

厦门大学

硕 士 学 位 论 文

公安地州级数据仓库的研究和设计

Research and Design of Police Information Data Warehouse

张秀成
指导教师姓名：史亮　副教授

专 业 名 称：软件工程

论文提交日期：2011 年 4 月

论文答辩时间：2011 年 5 月

学位授予日期：2011 年 6 月

答辩委员会主席：_____

评　　阅　　人：_____

2011 年 4 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

公安主要业务信息应用系统在全国范围内得到了推广，大多数城市的户政、治安、刑侦等业务部门建立了一大批应用系统，积累了大量的基础数据。由于历史原因，虽然各业务部门的单项应用得到了长足发展，但综合应用还比较薄弱，跨警种、跨部门的信息综合应用远未实现，信息共享程度较低，普遍存在信息孤岛现象。

某局在 2007 年建设了《综合信息平台》，整合各部门信息，采用较简单的技术方式已经收集整合了 10 多类公安业务数据，在建设公安综合信息应用和跨部门数据共享上发挥了重要作用。但当时仅考虑了结构性数据的整合、存储，没有考虑多种类型以及其它社会信息资源的整合、管理，也没有建立起可靠有效的数据资源来源整合、维护和对外服务的管理机制和技术体系，因此，数据质量和实时性以及信息资源类型已经严重不适应警务信息化发展对信息资源共享的需要。

本文工作主要针对当前某局在信息共享、资源优化方面的问题，旨在建设共享信息资源仓库，实现数据源提取、转换、整合、装载子系统；共享信息资源仓库分类数据库、资源统一维护、管理子系统；共享信息资源仓库对外数据服务管理子系统和提供集成应用服务，充分发挥信息资源的价值效能。

关键词：公安；数据仓库；信息共享

Abstract

Police main business information application systems have been promoted across the country. Most of the city's household registration, public order, criminal investigation departments have being established a large number of applications systems, accumulated a large amount of basic data, saving police resources effectively and enhance Police force's combat effectiveness. Some developed regions, the public security system network business applications have reached a certain size.

In 2007, Our bureaus established “integrated Information Platform”, integrated information of various departments, used relatively simple technical means which have collected more than 10-class police business data, which played an important role in police general information application construction and inter-departmental data sharing, and great convenience for the routine police work. However, due to objective conditions at that time , we just considered the structured data integration and storage, without considering various types of information resources and other social information integration, reliable and effective integration of data resources management, and didn’t establish management system and technology system maintenance and external service, therefore, that data quality and Real-time and information resources can not meet the police information need so has been a serious to the development of information sharing.

With the rapid development of IT technology, high-end cross-sectoral , inter-regional application of model-based application development, which require not only a simple extraction, collection, storage. By expanding the information resources (including structural data of social resources information resources, text information resources, audio information resources, graphics and image information resources, system information resources of different structure, Classified information resources etc.), rich source of information extraction channels and methods, improve the level of maintenance and management of information resources.

Keywords: Police; Data Warehouse; Sharing

目 录

Contents

第一章 绪 论

1.1 研究背景与意义

由于历史上的原因，虽然公安机关各业务部门的单项应用得到了长足发展，部分发达地区在一定程度上实现了某些有限信息的共享，但网络化应用、综合应用还相当薄弱，跨警种、跨部门的信息综合应用远未实现。从总体上看，信息共享程度较低，普遍存在信息孤岛现象。

某地区在 2007 年建设了“综合信息平台”，采用较简单的技术方式已经收集整合了 10 多类公安业务数据，在建设公安综合信息应用和跨部门数据共享上发挥了重要作用，极大方便了广大民警的日常警务工作。但是，由于当时的客观条件限制，仅考虑了结构性数据的整合、存储，没有考虑多种类型以及其它社会信息资源的整合、管理，也没有建立起可靠有效的数据资源整合、维护和对外服务的管理机制和技术体系，因此，数据质量的实时性以及信息资源类型已经严重不能适应警务信息化发展对信息资源共享的需要。

随着 IT 技术的飞速发展，以综合高端的跨部门、跨地区应用为主的应用模式发展方向，要求信息资源的整合共享不仅是简单的抽取、汇集、存储，其需要着重解决的是信息资源管理服务职能的充分体现和完善、优化，通过扩展信息资源类别（包括社会信息资源的结构性数据资源、文本信息资源、音频信息资源、图形、图像信息资源，异结构体系信息资源、涉密应用信息资源等）、丰富信息源提取渠道和方法、提高信息资源维护管理水平、增强信息资源多样性服务能力、建立长效集成服务机制等手段，并且建立一系列共享信息数据获取、整合、维护、服务等技术体系和维护管理机制，采用有效、可靠、安全、先进的技术措施使信息数据的汇集整合、分类整理、标准化规范处理、实时性校验、差错处理、向业务应用系统提供规范标准的共享信息服务接口等一系列重点、难点问题得到有效解决，按照“完整、准确、鲜活”的要求，将静态的信息资源转变为动态的信息资源。这样才能充分发挥信息资源的价值效能。

1.2 建设目标和任务

本项目的建设目标是：充分利用现有的网络、系统以及数据资源，扩充必要的软、硬件设施，建设地州级数据仓库，实现社会信息资源数据的汇集管理与服务。信息数据类型包括结构性数据资源、文本信息资源，异结构体系信息资源和涉密应用信息资源等多种信息资源。采用数据仓库建立不同类型的资源数据库和管理服务体系，依据“完整、准确、鲜活”的原则要求，依照不同信息资源对象建立不同类型的数据对象库和对象文件系统，以便于分类维护、管理，提高信息数据处理和对外服务效率及可管理性。同时应采取信息数据入、出库和分类处理的优化技术措施，以满足对海量数据和大数据对象的处理、交换和管理能力。按照我行业（GA/T543-2005，2011 修订版）和《综合共享数据库设计规范及应用》的相关标准和编制规范，建立数据标准化机制和统一的管理与对外共享服务机制，为综合、高端业务应用平台、业务部门应用系统及其它共享应用提供标准化、高质量的共享信息数据支持。

1.3 需要整合的数据

需要整合的数据有公安机关各业务系统数据以及民航、企业、人口计生基本信息、电信运营商用户信息、旅游人员信息、公路交通车辆及乘客信息企业信息等全社会能够整合的所有信息。

1.4 论文主要工作

本文工作的主要目的在于建设共享信息资源仓库，实现数据源提取、转换、整合、装载子系统；共享信息资源仓库分类数据库、资源统一维护、管理子系统；共享信息资源仓库对外数据服务管理子系统和提供集成应用服务。论文的主要研究内容如下：

1、 调研整理与公安工作相关的社会信息共享数据资源，按照《业务基础数据元素集》（GA/T543-2005，2011 修订版）等相关标准和编制规范形成全地区共享数据资源共享标准规范。

2、 通过数据资源调研，按照数据资源的不同类型，分别设计数据分类组织

模式、数据结构、数据关联关系、分类数据对象类型库建库模型等。

3、 依据分类数据库建库模型，准备并部署网络及硬件环境，建立地州级共享数据资源仓库模型及管理体系。

4、 根据已有信息资源的不同业务应用系统特点，有针对性地设计高效、可靠、安全、简洁、实用的共享信息提取、转换、整合、同步方式。

5、建立共享数据资源仓库维护管理子系统，实现共享数据资源库元数据、数据关联关系、数据质量、数据代码标准等内容的统一管理。

6、 实现共享数据资源仓库统一的对外共享信息服务。实现按需组织的专题数据库共享、数据接口共享等多种共享方式的配置与审计管理；逐步提供可与专题应用对接的数据服务代理和数据服务转发功能。

1.5 论文组织结构

本文后续章节的组织结构如下：

第二章 介绍数据仓库背景知识，数据仓库体系结构、ETL。

第三章 介绍地州级共享数据资源仓库模型

第四章 研究数据资源仓库的数据抽取过程。

第五章 介绍资源信息管理的设计方案。

第六章 总结全文，并展望下一步工作。

第二章　数据仓库相关知识介绍

数据仓库[1-4]是近年来兴起的一种新的数据库应用。在本章中，将就系统所涉及到的数据仓库、数据仓库的特点、数据仓库的设计和建立、数据仓库的建模以及 ELT 的定义和特点，操作流程等基本概念进行介绍。

2.1 数据仓库的概念

数据仓库是被誉为“数据仓库之父”William H.Inmon 的《建立数据仓库》一书中首次提出。随着人们对大型数据系统研究方面的深刻识认和不断完善，在总结和集中多行企业信息的经验之后，为数据仓库给出了更为精确的定义，即“数据仓库是在企业管理和决策中面向主题的、集成的、与时间相关的、不可修改的数据集合”。由于传统的数据库技术是面向应用的，通常针对不同的应用建立不同的数据库，因此，数据在不同的数据库中分散地存储，不易于统一查询，而且通常不保存历史数据轨迹，在这种情况下对数据进行分析，往往缺乏可靠性。数据仓库具有强烈的工程性，通常按其关键技术部份分为数据的抽取、存储与管理以及数据的表现等三个基本方面[5]。

数据仓库是决策支持系统和联机分析应用数据源的结构化数据环境。数据仓库研究和解决从数据库中获取信息的问题。数据仓库的特征在于面向主题、集成性、稳定性和时变性。

可能会把数据仓库简单地理解为仅仅是一个大型的数据存储机制，是一个静态的概念，实际上，数据仓库更像一个过程，这个过程涉及数据的收集、整理和加工，生成决策所需要的信息，并且最终把这些信息提供给需要这些信息的使用者，供他们做出正确决策。数据仓库的重点与要求就是能够准确、安全、可靠地从业务系统中取出数据，经过加工转换成有规律信息之后，供管理人员进行分析使用。因此数据仓库是一个动态的概念，应该称为数据仓库工程（Data Warehousing）。

2.2 数据仓库的特点

数据仓库是近年来兴起的一种新的数据库应用。虽然目前已开发的数据仓库数不胜数，但是大部分遵循着如下几个特点：

1、数据仓库是面向主题的；

2、数据仓库是集成的，数据仓库的数据有来自于分散的操作型数据，将所需数据从原来的数据中抽取出来，进行加工与集成，统一与综合之后才能进入数据仓库；

3、数据仓库是不可更新的，数据仓库主要是为决策分析提供数据，所涉及的操作主要是数据的查询；

4、数据仓库是随时间而变化的，传统的关系数据库系统比较适合处理格式化的数据，能够较好的满足商业商务处理的需求。稳定的数据以只读格式保存，且不随时间改变。

5、汇总的。操作性数据映射成决策可用的格式。

6、大容量。时间序列数据集合通常都非常大。

7、非规范化的。Dw 数据可以是而且经常是冗余的。

8、元数据。将描述数据的数据保存起来。

9、数据源。数据来自内部的和外部的非集成操作系统。

2.3 数据仓库的建立与设计

由于企业的各业务系统是不同时期、不同背景、不同开发商开发、面向不同的应用的，其数据结构、存储平台和系统环境均存在着很大的差异，不同的环境对于建立和设计数据仓库会带来不同的问题。

2.3.1 管理层的支持

数据仓库可以对分散在企业各处的数据进行整合和分析，从而帮助他们做出更好的决策，目前这方面已经有许多成功的应用[8-15]。在数据仓库应用程序实现以后，并不会对公司原有的操作进行改变。有时，通过建立一个概念证明模型(proof-of-a-concept， POC)可以帮助解释商业智能(BI)。POC 数据仓库只包含全

部数据的很小的一部分，并且只暴露给少数用户，以便进行测试。

2.3.2 数据有效性

数据仓库经常需要将许多的数据源进行连接，有时，这些数据源并不那么明显或者容易获得。根据企业政策的不同，很难得到所有必要的数据元素，有些数据可能包含机密信息或者高明感度的细节。

数据仓库是对现有系统的补充而不是替代。

2.3.3 复杂性

在的时候，数据可能会分散在数据库管理系统、电子表格、电子邮件系统、电子文档甚至纸上。要想办法解决如何数据仓库得到所需要的数据源，并为它们建立一个统一的模型。

2.3.4 数据质量

建立数据仓库是最大的问题在于如果应用程序没有检验数据的有效性，可能是遇到的字符串值很肯能使缩写的、拼写错误的或者完全忽略的。对于事务级的报告，这可能并不是一个严重的问题，将数据归类或者为用户提供据测能力帮助时，数据质量问题是非常严重的。

应用程序必须要反映出数据仓库内的数据源所发生的变化，特别是维度表中的数据，他们更容易改变。

首先，最简单的方法是重写现有的记录而不记录变化。

第二个管理维度变化的方法是在值发生变化时创建一个新的纪录，并将旧记录标记为废除。

第三个也是最后一个方法是在维度表的相同行不同列中维护历史值的变化空间。

2.4 数据仓库的建模

信息报表是维度建模的基础：三类数据实体：指标或度量单位，商业维度，商业维度的属性。、

通过信息报表构造了事实表和维度表。表在维度模型中如何安排？他们在模型中的关系如何？如何标记这些关系？查询和分析有那些类型？这些都是目前要重点解决的问题。通过多个维度表的维度属性，查询事实表中的指标就是典型的查询与分析。例：查询 2010 年版本的 Cherokee、在 2011 年 1 月份 ABC 经销商卖出的、客户信息、并通过银行提供的信息等。符合以上条件的交易有多少？需要通过多个维度表中的属性分析所有这些事实。一个查询中会用到每个商业维度表的部分或者全部属性。

在数据库仓库的建模中要分析前提，将事实表放在中央，维度表安排在事实表的四周能够满足这些要求。事实表位于星型中央，维度表分布在星型的各个角上——星型模式。构建星型模式是数据仓库建模的基本数据设计技术。维度表的内容:维度表的集合是星型模式中的关键部分。

星型模式是一种关系模型，非规范化的关系；维度表与事实表之间是 1：N 的关系；星型模型的优势为，用户容易理解；OLTP 用户使用预先定义好的 UI 或者查询语句与系统进行交互，用户不需了解数据模型；但 OLAP 是用户驱动的[6, 7]，用户必须清楚数据模型，星型模型容易被用户理解，完全按照与用户相同的理解关系的方式定义了连接路径[10]。（OLTP 中表的连接关系可能要穿越规范化的 N 个表才能了解到表之间的关系），星型模型这种优势不仅仅体现在后期使用、理解方面，在前期的开发阶段也便于和用户进行交流。在数据库模式中，表与表连接的目的在于寻找到需要的数据。如果连接的路径复杂，那么在数据库中浏览数据将是缓慢而艰难的。如果连接路径简单、直接，则浏览数据会更快。星型模型的优势之一在于它优化对数据库的浏览。

2.5 ETL

ETL 将数据抽取（Extract）、转换（Transform）、清洗（Cleansing）、装载（Load）的过程[1, 4]。它是构建数据仓库的重要环节。数据仓库是面向主题的、集成的、稳定的且随时间不断变化的数据集合，用以支持经营管理中的决策制定过程。数据仓库系统中有可能存在着大量的噪声数据,引起的主要原因有：滥用缩写词、惯用语、数据输入错误、重复记录、丢失值、拼写变化等。即便是一个设计和规划良好的数据库系统，如果其中存在着大量的噪声数据，那么这个系统也是没有任何意义的，系统根本就不可能为决策分析系统提供任何支持。为了清除噪声数