基于数据仓库的组织人事信息系统的设计实现

Probe on Organization Human Resource Data Analysis

窦忠秋　沈伯青　马 明　(上海市组织人事信息技术服务中心 201100)

摘要:随着计算机技术、网络技术和数据库技术的飞速发展,政府的信息化进程正在逐渐加快。在"党管人才"的大背景下,上海市委组织部利用积累二十多年的信息资源,采用数据仓库思想率先在信息分析和数据挖掘方面进行了探索和实际应用,有力地推动了党务部门信息化的进程。本文重点阐明了采用数据仓库思想,利用联机分析处理、数据清洗转换工具和数据分析展现工具为手段的一整套人事信息数据分析的开发思路和具体实现。通过该项目的实施,能够向各类工作人员提供个性化的多维信息,使分析处理信息的能力和信息的利用率大为提高,并帮助业务部门和领导更快更好地制定和做出决策,为传统的组织人事数据分析和管理提出了新的思路和方法,在业界有一定的借鉴作用和推广意义。

关键词:信息分析 组织人事 数据仓库 数据转换

1 引言

在近年政务信息化的发展过程中,组织部门作为党的人力资源管理机构,从八十年代起,以上海市领导干部多媒体管理信息系统为主要工作平台,逐步形成了"中央—市(省)—区(委办)—街道"四层应用体系。近年来,积累了全市局级干部、局级后备干部、处级干部以及全市党政人才信息库等不同层次的信息资源。我们根据业务需要开发了一系列应用系统,如领导干部收入申报信息系统、干部任免审批信息系统、干部出国政审信息系统和年报统计系统等,这在一定程度上提高了组织部门信息化的利用水平,但由于各业务系统互不关联,形成了一个个信息孤岛,给领导和业务部门的信息利用带来了很多不便。如何将积累了二十多年的管理信息系统中的数据得到充分利用? 在"党管人才"的大背景下,如何利用信息化来提高管理质量和管理水平? 如何贯彻"科教兴市"主战略,更科学地实施人才选拔、交流、培养和储备? 2004 年起,上海市委组织部以数据整合为契机,采用数据仓库理念,利用微软公司的 BizTalk 和海波龙公司的 Brio 等工具,探索建立领导干部决策支持信息分析系统。

两年来,上海市领导干部决策支持信息分析系统在组织系统得到了充分应用,在全市局级领导干部整体结构和发展趋势分析、上海市局级后备干部成长轨迹分析和区县处级干部功能定位配套分析等课题中发挥了重要作用,为领导决策和政策制订提供了客观的参考依据,得到了业务处室的广泛好评。本文先介绍基于数据仓库的信息分析系统的实现过程和体系结构,重点阐述实现中的难点和关键环节,并指出未来该系统进一步完善的方向。

2 信息分析系统的实施过程和体系结构

根据建立数据仓库的一般工作模式与步骤,结合组织部已有数据源的实际情况,上海市领导干部决策支持信息分析系统的实施按照"需求设计—>数据清洗与抽取转换—>前台展现系统的开发—>产生分析报告"进行。具体体系结构如图 1。

2.1 数据源

上海市领导干部信息管理系统已运行二十多年,数据源的格式主要有早期的 Foxpro 以及近年的 Access、Sql 等,从人员分类看,既有历年全市干部的统计汇总信息又有市管局级干部、市管局级后备干部、处级干部等不同层次的信息资源。从业务系统看,以市管

局级干部为例，有干部收入申报信息系统、干部任免审批信息系统、干部出国政审信息系统和年报系统等。

2.2 ECTL 过程

为了能对业务部门屏蔽复杂的业务系统，我们充分利用微软的 Biztalk 工具，并自行进行了本地化开发，将不同数据格式、不同业务系统的信息资源进行数据的抽取（Extract）、清洗（Clean）、转换（Transform）、以及加载（Load）等四个步骤，分别进行不同种类的维度可能是将姓名信息拆分为姓氏信息和名字信息，也可能是将基本维度信息进行再加工，例如学校维度转换，不仅仅是清洗和标识所有的学校名称、还需要依据相关规则来确定该学校是否属于国家重点院校、军队院校、党校等。加载即将转换后的数据加载到目标数据仓库的存储介质中。

2.3 数据存储管理

按照数据仓库的要求，我们进行了维度表、事实表

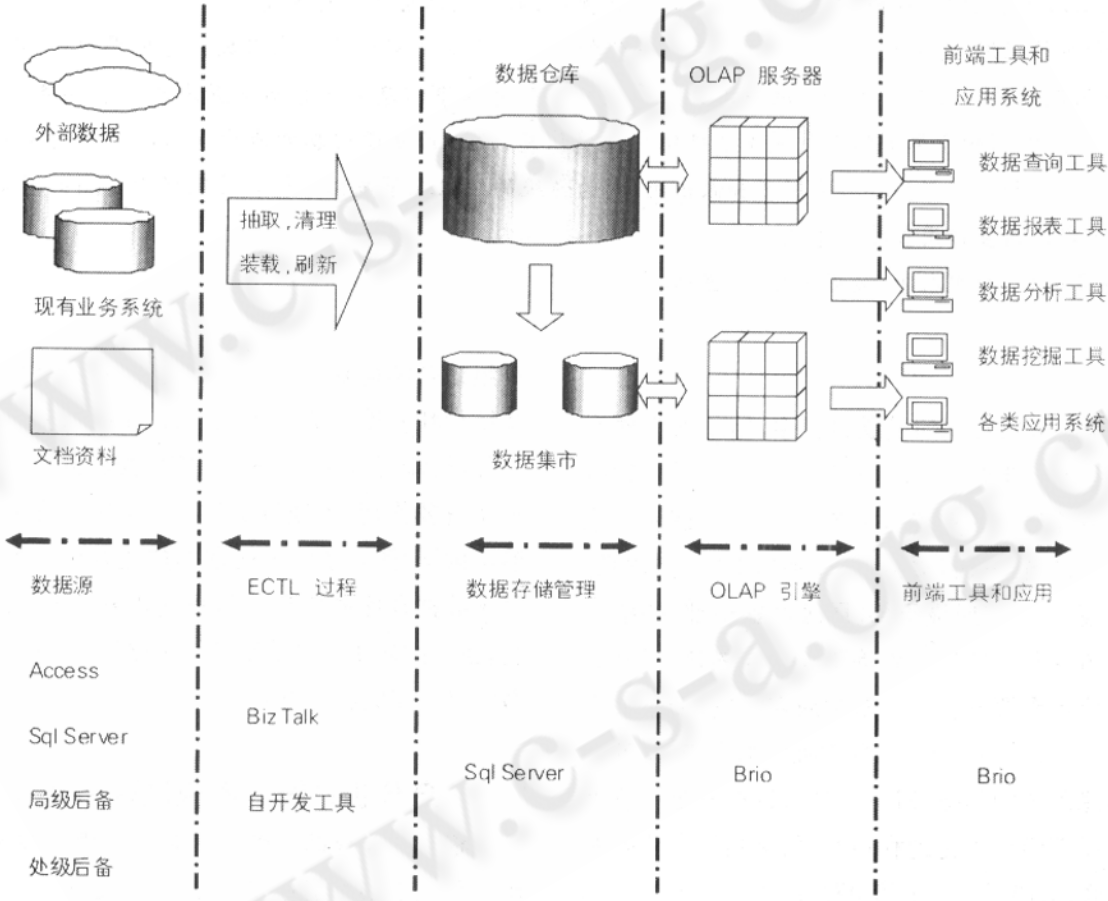

图 1 信息系统体系结构图

数据的转换。其中抽取的主要任务是确定业务数据的格式以及相关的业务数据源。业务数据往往需要进行清洗，因为业务数据在存储的过程中有时候并不规范，例如人员的毕业院校名称，上海交通大学在业务数据中存储的格式会有“交大、上交大、上海交大、交通大学”等，清洗的过程就是将这些不规范的数据进行标准化的过程。数据转换是将数据进行一定格式的转换，的设计，改变原有业务系统中数据的存储方式，专门为分析系统建立了特有的星型和雪花型模型存储结构，并将经过 ECTL 后的数据打碎后进行重新加载，统一存储到新的维度和事实表中，为下一步的 OLAP 作好了数据准备。

2.4 OLAP 引擎和前端展示工具

我们采用海波龙公司的 Brio 作为前台 B/S 展现工

具，这是因为它可以创建专有的数据存储格式文件并可以在 Web 上进行发布和数据钻取，同时 Brio 支持权限设定，这为专有数据的安全性保障带来了方便。通过 B/S 架构，用户可以仅仅依靠 Web 浏览器就可以实现对数据的访问、维度分析、数据钻取，同时 Brio 拥有的数据在线三维图表显示的功能也极大的促进了人机交互友好性。

3　信息分析系统实现的难点和关键环节

3.1　分析主题的确定

通过该项目的实施，我们认为确定分析主题需要具备三方面的知识——干部业务知识、计算机专业知识和数学建模知识，三者缺一不可。具体来说，干部业务人员往往从业务本身出发，提出一些可能有价值的分析主题，但计算机应用人员可能发现现有的支持数据不足，这样的分析很难得出相关的结论。有时计算机应用人员认为现有数据可以支持，但用聚类、递归等数学方法，数学模型始终确立不起来，也导致最终没有任何有价值的结论。因此，确定分析主题是制定后续信息分析的方向，是分析成败的关键。

对组织部门来说，分析主题的设置方向大致有三类：

（1）群体结构性分析，如领导干部总体结构分析，涉及年龄、学历、性别、民族等，或是具有某种特点的干部群体特征挖掘，涉及专长类型、专业背景、教育培训、成长轨迹等；

（2）对比类分析。即根据不同时间点、不同群体（班子）之间进行相关指标的对比分析。

（3）趋势预测分析。即在前两类型的基础上，通过大量信息和数学工具，预测未来发展趋势，提供必要的业务监测，为政策制定等提供辅助依据。

根据业务要求，结合分析主题的设置方向，可将分析主题分为常规分析主题和专题分析主题。

• 常规分析就是按一定周期（旬、半月或月）在格式相对固定的分析模板（根据工作需要，通常半年或一年更新一次模板）基础上作的业务分析。常规分析虽然也能产生类似于每月固定格式的统计报表，但其良好的互动和挖掘功能是统计报表所不能实现的，常规分析工作的目的是通过固定模式的分析判断其是否合理，从而发现问题并寻求引起问题的因素，最终提出解决问题的措施。

• 专题分析是根据组织系统业务的重点和热点问题，如今年围绕各级领导班子集中换届的班子对比分析，针对新经济组织的党员结构分析等都是结合一

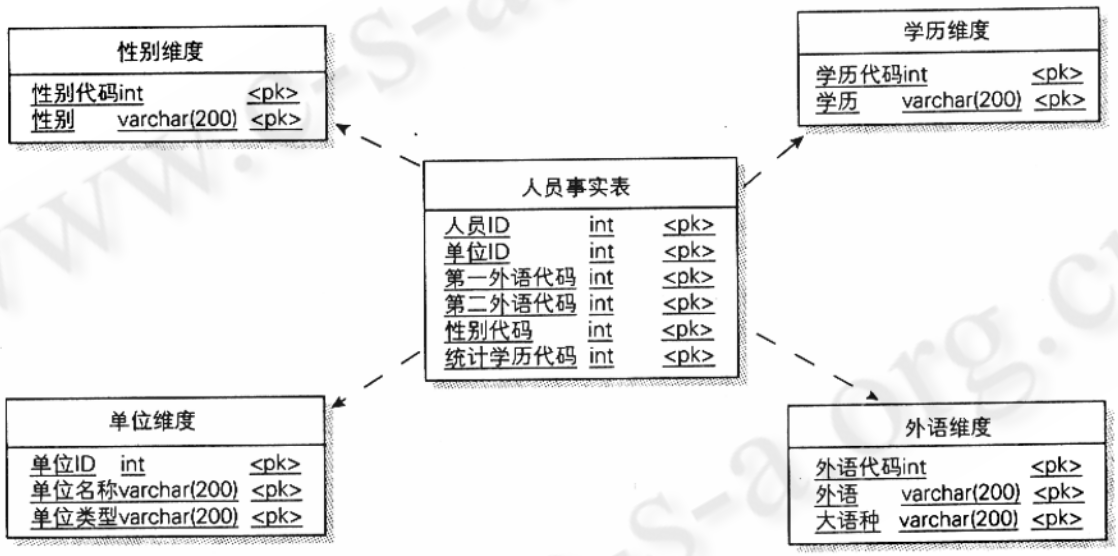

图 2　人员事实表存储结构示意图

段时期的工作重点和热点进行的。有时，我们还可以根据常规分析发现的异常情况而确定需要进一步深入进行相关的专题分析。专题分析没有固定的模板，因此每执行一次专题分析需要向数据仓库提交一次数据需求，包括多维度的数据表需求，给维度设计带来了一定的挑战。

3.2 数据维度的建立

建立数据维度是对分析主题完全理解的体现，它不但是整个业务实现的基础，也是数据仓库应用的核心。建立数据维度涉及数据仓库的理论和知识，它要求跳出传统管理信息系统的思维框架，面向分析主题，从业务角度确立相关指标，重新组织原有数据。如组织部门的分析主题都是与人相关的信息，如性别、年龄、民族等。传统的管理信息系统中人员基本表中往往含有如下信息：

表 1　人员基本字段表

单位	姓名	性别	民族	学历	外语程度

而在分析系统中，人员基本表常作为分析主题的事实表出现，存储形式如下。

在数据维度建立的过程中，关键要把握三点：

(1) 独立性是建立维度的基本原则

如性别维度、民族维度、健康维度等概念互不冲突，独立性强。性别维度的结构如下：

表 2　性别维度表

代码	描述
1	男性
2	女性

(2) 层次粒度是建立维度的重点

用户思维的逻辑惯性要求先看到概貌，再将关心的信息逐层深入展现，即逐层数据钻取，这正是面向主题进行分析的优势所在。要达到这种效果，一定在设计维度时要考虑到层次粒度概念。如由于管理信息系统中民族有近百种描述，在民族维度设计时，我们按照用户的思维习惯，往往先关注民族的分布概貌，即是汉族还是少数民族，其次对少数民族，想了解少数民族中人数较多的五大少数民族(藏、蒙、维、回、壮)等与其他少数民族的分布情况，最后才是每一个民族的分布情况。因此，设计维度时可采取如下结构：

表 3　民族维度表

代码	描述	是否少数民族	是否五大少数民族
01	汉族	0	0
02	蒙古族	1	1
03	回族	1	1
04	藏族	1	1
05	维吾尔族	1	1
06	苗族	1	0
07	彝族	1	0
08	壮族	1	1
09	布依族	1	0
10	朝鲜族	1	0
11	满族	1	0
……			
91	其他族	1	0

(3) 全面性是建立维度的难点

建立维度时，往往只考虑当前的分析主题，对未来可能出现的专题分析不可能考虑周到。针对专题分析建立新的数据维度是必要的，但最好避免，因为它会带来基础数据结构的变化，重新加载时可能会影响原有的分析主题。如外语维度，对常规分析来说，我们认为掌握两种外语(一外、二外)即可满足需求，对三外等往往忽略。设计结构如下：

表 4　外语维度表 1

第一外语名称	一外熟悉程度	第二外语名称	二外熟悉程度

但是，如果对掌握外语语种进行专题分析时，上述的设计就不能满足要求。因此，在设计时应留有一定的余量，尽可能考虑全面些，但又要避免余量过大带来的大量存储空间的浪费。因此，设计的难点在于要求

设计者必须既熟悉业务，又精通信息技术。综合分析后，对上述的外语维度设计如下：

表 5 外语维度表 2

第一外语名称	一外熟悉程度	第二外语名称	二外熟悉程度	第三外语名称	三外熟悉程度

4 展望

由于数据仓库是应用导向的系统，它立足于业务应用，而非单纯的技术。数据仓库项目的成功与否，很大程度上依赖于它的需求设计。因此该项目严格按照“业务和技术结合”的方式进行了充分的调研，近一半时间花在需求设计上，主要完成分析主题的设计和分析维度的建立。

项目建设之初，能对数据仓库成功演绎的开发商和用户都寥寥无几，开发商动辄要求购买价格不菲的软、硬件设备和分析工具等。我们在探索中坚持在科学、有效设计其功能的基础上，根据现有条件和需要，配置软、硬件设备、分析工具甚至数据挖掘工具，开发各类应用。本系统开发中我们充分利用原有设备，并选用可免费使用的微软产品 Biztalk 进行数据清洗和整合，只采购了性价比较好的前台数据展现工具，项目投资得到了有效控制。

我们的信息分析系统是在同类部门中最先完成的具有商业智能技术的业务系统，它促进了组织部业务部门和信息部门之间的工作交流，同时也在为领导在人才选拔决策上提供了良好的支持作用。同时，信息分析促进了用户维护相关管理信息系统的积极性和主动性，增强了信息分析系统与管理信息系统的互动，提高了管理信息系统的信息质量，形成了良好的信息反馈。将来我们将继续针对工作简历等非结构化信息分析的解决方案。

参考文献

1 数据仓库工具箱：维度建模的完全指南（第二版）[美]Ralph Kimball，Margy Ross 著，谭明金 译，出版社：电子工业出版社 ISBN:7505389319.

2 数据仓库生命周期工具箱：设计、开发和部署数据仓库的专家方法（美）Ralph Kimball Laura Reeves Margy Ross Warren Thornthwaite 肖明 王永红等译，电子工业出版社 ISBN:7505391925.

3 公共仓库元模型：数据仓库集成标准导论 [美]普尔等 著，彭蓉等译 机械工业出版社 ISBN:7111136020.