决策支持在海量数据系统中的实现*)

卢朝霞[1] 习 捷[1] 王 剑[2] 刘积仁[1]

(东北大学软件中心 沈阳 110004)[1] (东北大学信息科学与工程学院 沈阳 110004)[2]

摘 要 探讨了某海量数据系统中实现统计分析的策略和方法,并结合某大型人口信息系统中联机分析处理技术的具体应用,提出了在数据仓库模式下统计分析系统通用的功能架构。文章还针对实际情况,提出了合理的数据存储实现模式,并对在线分析系统的实现策略和指标库维度设计和优化过程进行了探讨。这种设计实现了对海量数据进行灵活、方便的查询和统计这一最终应用目标,将系统蕴含的基础数据转化为决策知识,也为大规模数据统计分析处理提供了一套完整的解决方案。

关键词 数据仓库,关系 OLAP,数据存储模式,维,统计分析

1 引言

信息技术的发展与广泛应用在系统中产生了大量的数据,这些数据包含了对生产和决策具有很高价值的信息[1]。如何将这些大量散乱的数据转化为决策知识,从而为相关部门提供决策数据支持及辅助决策分析,已成为亟待解决的问题。某大型人口信息系统中,包含了全国 13 亿人口的数据,其中的统计分析系统提供统计分析及决策支持服务以便为各相关部门更好地掌握与管理庞大的人口信息。目前系统的数据量已达到了 TB 级,在对如此一个海量数据系统进行统计分析的设计过程中,是否选用数据仓库,应该采取何种物理形式去存储这些统计数据,如何进行维度设计才能有效地提高统计的效率,以及应该采用何种构架策略,这些问题对在线统计分析的实现提出了挑战。

2 在线分析系统的体系结构和实现策略

某大型人口系统从人口基本信息和人口管理业务信息库中定期获取统计指标,加载到指标库当中,通过在线分析,提供对人口地域分布情况、人口迁移、姓氏用字等多种内容的统计和预测。总的来说,如果一个系统中存在历史数据和众多不同类型的数据源,就应采取建立数据仓库的方式,系统从这些数据源抽取数据并进行清理、整合、汇集和再组织,发挥数据汇集优势,对这些信息进行整理和分析挖掘[2]。一般情况下,这样的数据仓库系统的架构模式如图 1 所示。

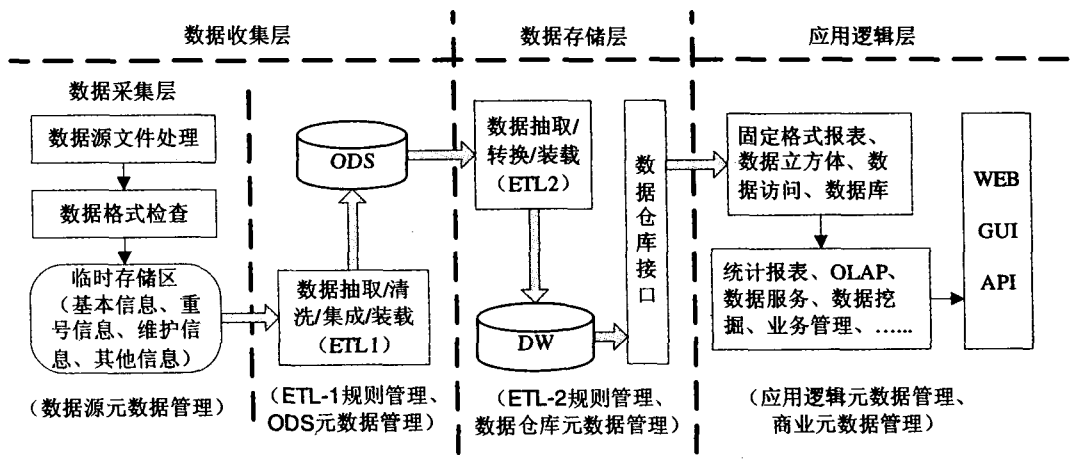

图 1 统计分析系统功能结构图

这样的设计中,数据收集层按照统一数据源接口标准,从各个专用业务信息库中进行抽取、转换、装载,形成数据仓库所需的信息;数据存储层对数据仓库中各级信息进行组织、存储与访问控制管理;数

*)基金项目:国家 863 计划资助课题"企业应用集成平台及集成构件的开发与应用"(编号:2004AA1Z2370)。卢朝霞 教授;习 捷 博士研究生,研究方向为数据库与信息安全技术的应用。

据展现层使得用户和有关系统通过上层提供的各种应用，可以利用 Web、GUI、API 等方式对人口信息进行查看和访问。其中，ODS(操作数据存储)是面向主题的(Subject -Oriented)、集成的、可变的、当前或接近当前的数据集合[3]。从某种角度来说，它实际上就是数据准备层，在建立物理级数据仓库系统 ETL 前先对准备抽取的数据作一次准备，从而避免无效和不安全的抽取数据。这种方法的优点是可以加快在线数据查询速度，避免无效数据和错误数据的抽取，保证数据仓库的完整性和一致性[4]。在一般的数据仓库系统中，利用如上的构架，即可合理地组织数据的存储和使用，使对数据进行多角度的统计分析和提供决策支持成为可能。

以上是常规情况下系统的建设策略，一般来说，在存在众多不同类型的数据源和存在大量的历史数据的情况下，必须建立数据仓库从这些数据源抽取数据并进行清理、整合、汇集和再组织，对这些数据进行整理、分析挖掘。但是某大型人口信息管理系统本身就是各省、市人口系统数据的汇集，数据量虽然非常大，但有着数据结构单一的特点，并且数据在数据库中长时间存放，可以实现信息追溯。所以在未来较长时间内，仅就人口信息管理系统本身而言，可以不必建立数据仓库系统。条件成熟后，可以参照上述构架，用以下两种方式建立数据仓库系统：在人口信息管理系统的历史数据逐渐增多后，建立数据仓库；或者与其他全国信息资源库共同建立数据仓库。所以，在建立数据仓库系统之前，可以首先建立在线分析系统，直接面向人口信息管理系统获取数据，完成统计分析功能。

3 系统指标库的数据建模方法

统计分析指标库包括基本信息指标库和业务信息维护指标库，预计这些数据所占的空间在 2006 年底将达到 1.54T，2007 年会到达 2.05T，2008 年会超过 2.56T。对于已经达到 TB 数据量级的系统，为了保证系统的可用性和有效性，首先必须采用一种合理的数据建模方法，才能保障整个系统的效率。

综合考虑系统的数据量和对性能的要求，为了尽可能减少对生产业务系统的影响，指标库的设计采用 ROLAP 方式实现对信息的统计分析。ROLAP 是基于关系数据库的 OLAP 实现(Relational OLAP)，它以关系数据库为核心，以关系型结构进行多维数据的表示和存储[5]。ROLAP 可以节省存储空间，有很大的灵活性，并且与关系数据库保持一致，有着明显的优势[6]；而 MOLAP 虽然在性能和管理的简便、查询速度比较快，但是也有加载、维护以及存储结构和空间等方面的限制[7]。所以，在系统建设的初期，考虑到数据比较单一并且数据在库中长期存放，最终采用了 ROLAP 方式来构建星型模型为主的方式。

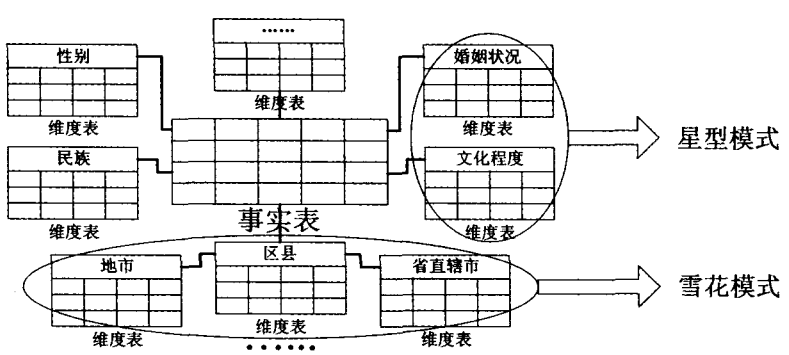

图 2 指标库设计示意图

维表和事实表通过主关键字和外关键字联系在一起，形成了“星型模式”。对于层次复杂的维，为避免冗余数据占用过大的存储空间，可以使用多个表来描述，这种星型模式的扩展称为“雪花模式”[8]。人口信息系统中，根据区县维度表、地市维度表、省市维度表与事实表之间复杂的关系，这部分的设计采用了“雪花模式”，图 2 是指标库设计的示意图。通过这种方式，在生产库基础上建立统计分析指标库，只需定期生成指标库，以后的统计分析主要都在指标库上进行。

4 指标库维度的设计过程

有了好的数据存储方式和数据建模，指标库维度的设计也很重要，如果维度设计不当，就会导致查询统计有严重的效率问题。而在指标库维度设计中，每个维度有其粒度和属性值，由于指标库的最大数据量即所有维度属性值的笛卡尔积，所以控制维度的数量、控制其粒度的大小和属性的数量，是保证统计的效率的关键之一。指标库的设计按照统计分析所关注和考察的角度进行设计，在控制指标库的数据量上，遵循只取主要的业务维度和减少维度粒

度的原则。指标库的设计包括基本信息指标库、姓氏分析指标库、业务信息指标库等，现以基本信息指标库为例，说明指标库的设计过程。由于其中包含一些敏感信息，关于具体维度的名称，此处不再一一描述。表1中描述了指标库维度设计的过程。

表1 维度设计过程表

	初始设计	第一次优化	第二次优化
设计策略/优化方法	16个维度，适当划分粒度	1.对指标库每个维度创建BITMAP索引； 2.对维度表与多维指标库之间建立关联的BITMAP JION索引； 3.数据库设置里，打开数据仓库星形模式开关，将参数STAR_TRANSFORMATION_ENABLE调整为TRUE，该参数可以自动优化查询，将维表关联好再关联指标库的事实表； 4.调整PGA_AGGERATE_TARGET的大小，提高数据库对大数据量的汇总效率； 5.统计分析语句采用ORACLE HINT处理，执行计划使用BITMAP JION索引。	通过控制指标库的维度，缩减指标库的数据量来控制效率： 1.缩减维度，去掉不必要的维度； 2.减少维度的粒度。 最后缩减到10个维度，去掉身高、兵役状况、信息级别等8个维度，但是增加注销标识和历史维度2个维度。
指标库最大数据量	Cartesian product=∏[1≦i≦16]DimensionAttr(i) ≈3600*3*58*37*120*240*7*3600*10000*240*3600*7*100*3600*1000 =3.6623579161559E35>>全国人口数≈13E8		Cartesian product =∏[1≦i≦10]DimensionAttr(i) ≈378*3*58*10*5*120*340*34*8 =3.649556736E13
测试结果	基本信息数据量为5.4亿时抽取指标库数据量为5亿。两维组合分析响应时间为1小时；而三维及三维以上组合分析响应时间在2小时以上。	统计效率提高了一半左右，但至少还需要30分钟以上，效率依然不高，没有根本解决问题。且所选的分析工具只能执行自己生成的语句，无法通过手工加入ORACLE HINT处理使用BIT MAP索引。	基本信息数据量为6.5亿情况下，抽取的指标库为1千万，对多个维度组合的统计效率在5-12分钟左右，结果可接受。

在优化前，实际指标库的数据量十分接近原始数据量的值。第一次优化只是从数据库设置方面进行调优，但是没有从根本上解决问题。而在第二次优化中，是从关键出发，通过减少维度的粒度来提高效率。举例来说，优化前地域维度的粒度为“区县”，属性值为3600；优化后粒度为“地市”，属性值缩减到378，这样可以大大减少指标库的理论最大数据量，从而提高系统效率。通过以上的设计和优化，指标库的数据量比最初的设计大大减少，实际测试过程中的数据也表明，系统的效率也完全达到了可以接受的结果。而且，即使增加时间维度，依据历史维度的频度为一个季度或半年，所以其增加的数据量至少在3～5年内还是一个可以接受的范围，而这段时间以后，可以采用将旧的存档指标库进行再缩减的方法，或者直接采用数据仓库模式来解决。

结束语 通过以上的设计策略，在实际的运行环境中，系统已经达到指标库抽取时间2.3小时左右、结果展现5～10分钟的性能。统计分析系统在并发访问能力方面支持30个同时访问用户，在处理效率方面可以达到全国各类统计表单张形成时间小于5小时，进行切片、钻取等在线分析时，单次点击反应时间小于1小时，这些数据充分说明了此种设计策略的可用性。在这种TB数据量级的情况下，文章所提出的在线分析系统的实现策略和物理组织，以及指标库的维度的设计优化方案，可以对已入库的数据进行多角度的统计分析，为相关部门提供相应的统计分析数据支持，从而有效地提升了整体决策水平。

参 考 文 献

1 曹斌，齐剑锋，涂序彦. 商业智能决策支持系统(BIDSS)[J]. 计算机工程与应用，2001(20)：29～31

2 Inmon W H. Building the Data Warehouse (Third Edition)[M]. John Wiley & Sons，2002.20～53

3 王元珍，李海波. 基于OLE DB的数据抽取、转换和装入工具的设计与实现[J]. 小型微型计算机系统，2002，23(1)：453～455

4 王志谦，朱长征，陈福集. 数据预处理在商业企业数据仓库的应用[J]. 合肥工业大学学报(自然科学版)，2002，25(2)：286～289

5 Cody W F，Kreulen J T，Krishna V，Spangler W S. The Integration of Business Intelligence and Knowledge Management[J]. IBM Systems Journal，2002，41(4)：697～714

6 Han Jiawei，Kamber M. Data Mining，Concepts and Techniques[M]. Morgan Kaufmann Publishers，Inc，2001.39～41

7 Vassiliadis P，Simitsis A，Skiadopoulos S. Conceptual modeling for ETL Processes[M]. ACM Press，2002

8 张宁，贾自艳，史忠植. 数据仓库中ETL技术的研究[J]. 计算机工程与应用，2002，38(24)：213～216