

数据仓库的实现技术

Implementation Techniques of Data Warehouse System

张宜红 樊惠娟 王能斌

(东南大学计算机系 南京 210096)

摘 要 Data Warehouse System(DWS) is intended to be used to manage the data needed for Decision Support System(DSS). This paper introduces the background of DWS, analyzes their data and query characteristics, and tentatively discusses its implementation techniques.

关键词 Data Warehouse System, Materialized views, Decision Support System

70年代起,为满足现代管理的需要,人们在管理信息系统的基础上发展了以数据分析和建模定量分析为基础的决策支持系统,向决策者提供决策所需的信息,其核心部分是大量被良好管理着的数据。传统的数据库技术在数据共享、数据与应用程序的独立性、维护数据的一致性和完整性及数据的安全保密等方面提供了有效的手段,但仍存在以下问题:

首先,决策支持系统为掌握充分的信息,需要访问大量的数据,这些数据来源于不同的数据源,它们在定义、组织方式上不同,决策者为获取有关的数据需要熟悉不同的系统环境和数据定义,花费大量的时间理解数据的语义,一方面对决策者本身的素质要求过高,另一方面也不利于数据交流和信息提取。

其次,数据库中大量的事务性的、操作用的数据,即对每一项工作、管理对象的具体、细节性的描述。决策支持系统需要的是综合、总结性数据,特点是涉及数据量大,但不需了解其详细内容和个体的数据,要求对具体数据的概括,提取有用的信息。决策者不可能也不应花费时间浏览所有的细节数据。

最后一个是数据库中保存管理的是当前数据,而决策支持系统还要求有大量的档案数据,即历史数据以便于分析和比较,找出变化的趋势。

为了弥补数据库系统存在的以上不足,一种新的技术—数据仓库技术应运而生,成为研究数据管理技术的新课题。

一、基本框架

数据仓库是从多个内容相关,物理和逻辑上都

相互独立的数据源中提取的面向主题的数据集合。从数据库技术的角度看,数据仓库提供了多数据源上的一个统一的数据平台。图1给出了一个典型的数据仓库系统结构^[1]。

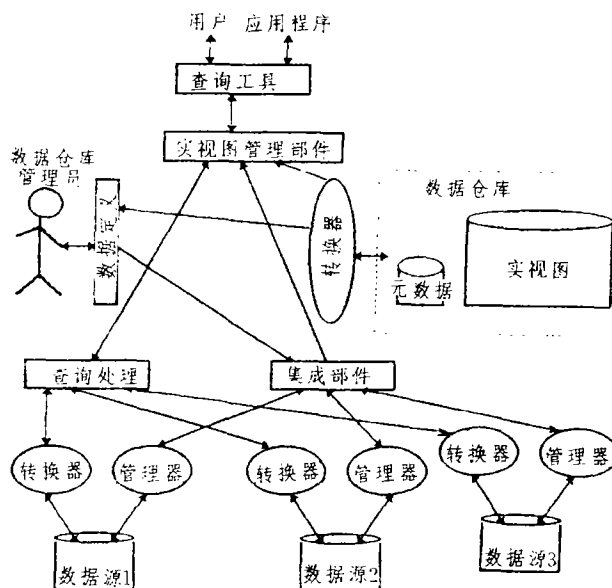


图1 一个数据仓库系统的典型结构

整个数据仓库系统由查询工具、管理部件、数据仓库、集成部件、查询处理和数据仓库管理员组成,数据仓库包括实视图和元数据。实视图是由其它数据源中的数据倒出。集成部件根据元数据中的有关定义,完成数据源中数据到实视图的变换,并负责获取源数据的修改内容。查询处理实现对源数据的查询。实视图管理部件负责实现数据共享、安全保密、实视图的维护和元数据的管理等工作。仓库管理员要对实视图管理部件、集成部件和查询处理部件进

行维护,定义并管理数据仓库中的数据。查询工具提供查询语言及可视化查询界面。以上几个部分互相协调,确保系统能快速、正确地提供用户所关心的数据。

二、实视图的性质

数据仓库系统中存放的是实视图,与一般的数据相比,具有以下特点:

- 时间是数据仓库中所有数据必备的属性。若采用关系模型,时间应是关系数据的主键组成部分,在同一个数据仓库中可以有同一对象在不同时期的状态描述,保存了它的历史数据。在开始创建数据仓库时,如果实视图所对应的源数据没有时间特性,在装载时还应加上相应的时间值。

- 数据一旦装入数据仓库中,基本不会发生变化,本文称为永久数据,这是由数据的时间性决定的。数据仓库中的每一数据项对应于某一特定时间。当对象某些属性发生变化,则生成新的数据项。这就使得数据具有稳定性,除非有大调整,数据第一次的存储位置就是它的永久存放位置;此外,数据仓库的数据增长很快,经过一段时间后,若存储器空间不够,就要进行适当调整。例如,把时间久远些的数据拷贝到其它存储器上,或对这些数据进行进一步的抽象(总结与综合),腾出空间装载新的数据。

- 数据仓库中的实视图不是数据源中数据的简单拷贝,而是经过包括总结、综合在内的一系列变换后的数据。综合可以是对同时间同类数据的综合,或是一段时间内数据的综合,对综合后的数据还可以再次总结、综合,一直循环下去,构成一个数据概括层次。从较为具体的到综合性强的数据叫数据的抽象,反之,称为数据的具体化。一般用户对数据仓库的查询沿抽象或具体化的方向进行。

- 为给决策者提供充分的信息,数据仓库中数据种类多,数据量十分庞大,可以超过万亿字节,存储在多个存储介质上,有时还需三级存储。

三、查询特点

数据仓库的查询特点是由其应用环境决定的。在决策支持应用环境下,一开始用户的查询目标并不十分明确,先取得总结性强的数据,获取一个总体概括,然后不断沿具体化方向查询,每次查询的结果决定下一次查询内容,属于交互式查询。这就要求系统的性能很强,能迅速以可视化的方式返回查找的内容。同时为给用户一个较好的查询导向,使用户尽

快地获取最终有价值的信息,系统应提供数据仓库中数据的特性描述,即对用户开放部分元数据。这包括按一定顺序排列的实视图定义,实视图层次关系,实视图建立的原因以及有关实视图内容的说明等诸如此类对用户查询有一定帮助的描述信息。

查询数据仓库的条件表达式一般比较复杂,大多要对实视图进行聚集函数运算,以方便在总体上进行比较、分析。相应地,数据仓库对聚集函数运算要有较强的优化处理能力。此外,数据仓库的查询条件必须有时间的约束,这也是数据仓库系统中的查询要复杂得多的原因之一。

四、数据仓库技术初探

前面已多次提到数据仓库中数据量十分庞大,对这么大的数据实现快速的查询响应,系统和管理上要有相应的技术支持^[2]。实视图就是预先生成供数据仓库查询的数据,这是提高查询速度的基本手段之一。除此之外,系统还需考虑其它技术,如:大缓冲技术、并行处理技术、查询优化、有效的存储管理、多维数据组织等。对实视图还应有相应的实视图维护技术。

1)内存管理^[3] 在数据仓库中,一般以读为主,数据读入内存后,一般不须写回磁盘。用户为某一次决策分析而进行的一系列查询之间可能存在语义上的关联,并呈现出逐步具体化的趋势,如果采用有效的缓冲调度算法,可以显著减少 I/O。

2)数据卷管理 过多并且不断增长的数据必然会产生多个数据卷。为完成对多个数据卷的管理,系统应能在不同存储介质之间实现数据转移,并能从象磁带、光盘这样的存储介质上读取数据。

3)并行处理技术^[4] 在数据仓库系统中为提高查询速度和减少实视图的生成和更新时间,必须采用并行处理技术。它包括数据的并行装载,同一查询的并行执行、查询与查询的并行和查询与数据装载之间的并行。

4)数据集成 在创建数据仓库时,数据集成的的工作包括格式转换、根据选择逻辑消除数据冲突、运算、总结、综合、统计、加时间属性和设置缺省值等工作;在实视图维护阶段,把源数据的变化过滤后以统一的格式通知实视图管理系统。集成器一般采用预编程方法实现,在初期整理数据源中数据的阶段,可适当使用辅助工具,帮助判别不同数据是否具有同一语义,减轻工作的繁琐程度^[5]。

5)实视图的维护 存在两个问题需要解决:一

是如何得知数据源中数据有变化;另一个是如何保持实视图与源数据的一致性。这两个问题都是数据仓库技术的研究热点。

到目前为止,对前一问题有三种解决方案。第一种直接从源系统获取修改信息,这包括一些有触发机制的主动数据库系统,能查询日志的系统和有增量数据的应用系统,但鉴于一些老的系统不具备以上条件,或不允许用户查看日志,这种方案不适用所有情况。第二种是修改源系统的应用程序,缺点是工作量大,而且通常不具有修改权限。最后一种是快照对比技术^[6],通过对比数据源的新旧版本,找区别点,自行获取修改信息,这一方案增加了实现一个数据仓库系统的工作量,而且在以后的工作中还要付出很高的时间与空间代价,是“最后的选择”。

对于第二个问题,现已达成基本共识:一般不采用重新生成的办法,考虑如何有效地实现增量维护法^[7],普遍的做法是根据实视图定义表达式和数据源的修改,生成实视图的修改表达式或实视图的维护步骤,最后完成实视图的维护。

6)时态数据的表示与处理^[8] 具有时间属性的数据被称为时态数据。从这一点上说,数据仓库的数据比一般数据库的数据要复杂得多。时态数据的时间值有两种语义:时间点与时间段,在数据仓库中,必须都能恰当地表示和处理。

7)安全保密^[9] 作为决策支持系统基础的数据仓库系统,由于它管理的数据往往是涉及全局的敏感数据,所以数据仓库系统应比一般的数据库系统有更可靠的安全保密功能。它必须保证在从数据源获取所需数据的同时,阻止恶意代码的侵入,防止机密数据的泄漏,避免遭受突破安全控制的攻击。

8)多维数据组织^[10] 多维数据是决策需要的一类特殊数据,它针对不同的查询有不同的数据抽象层次,所有的抽象层次构成数据立方。作为管理决策数据的数据仓库,必须有专门的组织结构支持对多维数据的查询。目前,多维数据库系统和R树都是专为多维数据服务的技术。

以上是数据仓库系统中的一些关键技术。要真正发挥数据仓库系统的作用,关键还在于查询响应时间、数据语义集成和实视图管理这三方面性能有质的飞跃,它们从根本上依赖于硬件性能提高、标准化的高效的实视图维护算法。在当前,如果能在数据仓库物理设计上充分发挥静态数据的优势,设计出比较完善的存储结构,很可能成为系统性能提高的一个突破点。

参考文献

- [1] J. Wiener et al., A system prototype for warehouse view maintenance, In Workshop on Materialized Views, Tech. and App., 1996
- [2] P. Fernandez and P. Schneider, The Ins and Outs (and Everything in Between) of Data Warehousing, In Proc. of ACM SIGMOD Intl. Conf. on Management of Data, 1996
- [3] P. Schenermann et al., A Data Warehouse Intelligent Cache Manager, Proc. of the 22nd VLDB Conf. 1996
- [4] M. N. Garofalakis and Y. E. Ioannidis, Multi-dimensional Resource Scheduling for Parallel Queries, Same to [2]
- [5] Wen-Syan Li and C. Clifton, Semantic Integration in Heterogeneous Databases Using Neural Networks, Proc. of the 20th VLDB Conf., 1994
- [6] W. J. Labio, and H. Garcia-Molina, Efficient Snapshot Differential Algorithms for Data Warehousing, Same to [3]
- [7] Y. Zhuge et al., View Maintenance in a Data Warehousing Environment, In Proc. of the ACM SIGMOD Intl. Conf. on Management of Data, 1995
- [8] D. Gabbay and P. McBrien, Temporal Logic and Historical Databases, In Proc. 17th Intl. Conf. On VLDB, 1990
- [9] K. S. Candan et al., Secure Mediated Databases, 12th Intl. Conf. on Data Eng., 1996
- [10] J. Gray et al., Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals, Same to [9]