数据仓库技术及实施

赵　方
(浙江树人大学,浙江 杭州 310015)

摘要:介绍了数据仓库的基本概念,针对数据仓库建立对创建数据仓库的过程进行了分析,对实现数据抽取、数据仓库的存储和管理等进行分析和比较。

关键词:数据仓库;联机分析处理;数据抽取;数据存储
中图分类号:TP311　　文献标识码:A　　文章编号:1009-3044(2006)17-0032-02

Research of Data Warehouse Technology

ZHAO Fang
(Zhejiang Shuren University,Hangzhou 310015,China)

Abstract:In this paper, the internal characteristics of DataWarehouse are introduced.Analyzed the procedure of integrated Data Warehouse and building the data warehouse, Data Extract,Data Warehouse Storage and how to manage the Data Warehouse.

Key words:Data Warehouse;OLAP(On-line Analytical Processing);Data Extract Transform Load;Data Storage

1 引言

传统的数据库技术是以单一的数据资源,即数据库为中心,进行事务处理、批处理、决策分析等各种数据处理工作,数据处理可划分为两大类:操作型处理(OLTP)和分析型处理(统计分析)。操作型处理也叫事务处理,是指对数据库联机的日常操作,通常是对一个或一组纪录的查询和修改,主要为企业的特定应用服务的,注重响应时间,数据的安全性和完整性;分析型处理则用于管理人员的决策分析,经常要访问大量的历史数据。而传统数据库系统利于应用的日常事务处理工作,而难于实现对数据分析处理要求,更无法满足数据处理多样化的要求。因此,专门为业务的统计分析建立一个数据中心,它是一个联机的系统,专门为分析统计和决策支持应用服务的,通过它可以满足决策支持和联机分析应用所要求的一切。这个数据中心就叫做数据仓库。

2 数据仓库概念及发展

2.1 什么是数据仓库

数据仓库就是面向主题的、集成的、不可更新的(稳定性)、随时间不断变化(不同时间)的数据集合,用以支持经营管理中的决策制定过程。数据仓库最根本的特点是物理地存放数据,而且这些数据并不是最新的、专有的,而是来源于其它数据库的。数据仓库的建立并不是要取代数据库,它要建立在一个较全面和完善的信息应用的基础上,用于支持高层决策分析,而事务处理数据库在企业的信息环境中承担的是日常操作性的任务。

2.2 相关基本概念

2.2.1 元数据

元数据(metadata):是"关于数据的数据",相当于数据库系统中的数据字典,指明了数据仓库中信息的内容和位置,刻画了数据的抽取和转换规则,存储了与数据仓库主题有关的各种信息,而且整个数据仓库的运行都是基于元数据的,如修改跟踪数据、抽取调度数据、同步捕获历史数据等。

2.2.2 OLAP(联机分析处理 On-line Analytical Processing)

数据仓库用于存储和管理面向决策主题的数据,OLAP对数据仓库中的数据分析,并将其转换成辅助决策信息。OLAP的一个重要特点是多维数据分析,这与数据仓库的多维数据组织正好形成相互结合、相互补充的关系。OLAP技术中比较典型的应用是对多维数据的切片和切块、钻取、旋转等,它便于使用者从不同角度提取有关数据,其基本思想是:企业的决策者应能灵活地操纵企业的数据,以多维的形式从多方面和多角度来观察企业的状态、了解企业的变化。对OLAP进行分类,按照存储方式的不同,可将OLAP分成ROLAP、MOLAP和HOLAP;ROLAP没有大小限制;现有的关系数据库的技术可以沿用;可以通过SQL实现详细数据与概要数据的储存;现有关系型数据库已经对OLAP做了很多优化,包括并行存储、并行查询、并行数据管理、基于成本的查询优化、位图索引、SQl的OLAP扩展等大大提高了ROALP的速度;可以针对SMP或MPP的结构进行查询优化。一般比MDD响应速度慢;只读、不支持有关预算的读写操作;SQL无法完成部分计算,主要是无法完成多行的计算,无法完成维之间的计算。

MOLAP性能好、响应速度快;专为OLAP所设计;支持高性能的决策支持计算;复杂的跨维计算;多用户的读写操作;行级的计算。增加系统复杂度,增加系统培训与维护费用;受操作系统平台中文件大小的限制,难以达到TB级;需要进行预计算,可能导致数据爆炸;无法支持维的动态变化;缺乏数据模型和数据访问的标准。

HOLAP综合了ROLAP和MOLAP的优点。它将常用的数据存储为MOLAP,不常用或临时的数据存储为ROLAP,这样就兼顾了ROLAP的伸缩性和MOLAP的灵活、纯粹的特点。

	优势	劣势
ROLAP	没有大小限制,现有的关系数据库的技术可以沿用;可以通过SQL实现详细数据与概要数据的储存,现有关系型数据库已经对OLAP做了很多优化,包括并行存储、并行查询、并行数据管理、基于成本的查询优化、位图索引、SQl的OLAP扩展等大大提高了ROALP的速度;可以针对SMP或MPP的结构进行查询优化。	一般比MDD响应速度慢;只读、不支持有关预算的读写操作;SQL无法完成部分计算,主要是无法完成多行的计算,无法完成维之间的计算。
MOLAP	性能好、响应速度快;专为OLAP所设计;支持高性能的决策支持计算;复杂的跨维计算;多用户的读写操作;行级的计算。	增加系统复杂度,增加系统培训与维护费用;受操作系统平台中文件大小的限制,难以达到TB级;需要进行预计算,可能导致数据爆炸;无法支持维的动态变化;缺乏数据模型和数据访问的标准。
HOLAP	综合了ROLAP和MOLAP的优点。它将常用的数据存储为MOLAP,不常用或临时的数据存储为ROLAP,这样就兼顾了ROLAP的伸缩性和MOLAP的灵活、纯粹的特点。	

收稿日期:2006-03-24

作者简介:赵方(1979-),女,浙江杭州人,浙江树人大学助教,硕士在读,主要从事教学、科研工作,以数据库应用、信息管理为主要研究方向。

2.2.3 粒度

数据仓库的数据单位中保存数据的细化或综合程度的级别。细化程度越高,粒度级就越小;相反,细化程度越低,粒度级就越大。粒度问题它深深地影响存放在数据仓库中的数据量的大小,同时影响数据仓库所能回答的查询类型。在数据仓库中的数据量大小与查询的详细程度之间要作出权衡。

2.2.4 样本数据库

一种粒度形式,即样本数据库。它根据给定的采样率从细节数据库中抽取出一个子集。这样样本数据库中的粒度就不是根据综合程度的不同来划分的,而是有采样率的高低来划分,采样粒度不同的样本数据库可以具有相同的数据综合程度。

2.2.5 分割

将数据分散到各自的物理单元中去,以便能分别独立处理,其目的在于提高效率;有许多数据分割的标准可供参考:如日期、地域、业务领域等等,也可以是其组合。一般而言,分割标准总应包括日期项,它十分自然而且分割均匀。

3 构建数据仓库

3.1 构建数据仓库的目标

数据仓库通过构造一种体系化的数据存贮环境,将分析决策所需的大量数据从传统的操作环境中分离出来,使分散的、不一致的操作数据转换成集成的、统一的信息,并通过不同纬度的分析、展示和钻取,和各种内部、外部数据的有效集成,为企业决策提供依据,从中快速找到对企业进一步发展有价值的潜在信息。

3.2 数据仓库的构建

数据仓库的构架由三部分组成:数据源、数据源转换/装载形成新数据库、OLAP。

数据仓库的实施过程大体可分为三个阶段:数据仓库的项目规划、设计和实施、维护调整。在技术上可以根据它的工作过程分为:数据抽取、存储和管理、数据的表现以及数据仓库的设计的技术咨询四个方面。

3.2.1 数据抽取

数据的抽取是数据进入仓库的入口。由于数据仓库是一个独立的数据环境, 它需要通过抽取过程将数据从联机事务处理系统、外部数据源、脱机的数据存储介质中导入到数据仓库。

数据仓库的数据不要求与联机事务处理系统保持实时的同步,因此数据抽取可以定时进行,但多个抽取操作执行的时间、相互的顺序、成败对数据仓库中信息的有效性则至关重要。

3.2.2 数据的存储和管理

数据的存储和管理是数据仓库的关键。数据仓库的组织管理方式决定了它有别于传统数据库的特性,同时也决定了其对外部数据表现形式。要决定采用什么产品和技术来建立数据仓库核心,则需要从数据仓库的技术特点着手分析。数据仓库的主要技术特征是对大量数据的存储和管理、并行处理能力、决策支持查询的优化及支持多维分析的查询模式,采用"星型模式"来组织数据的面向决策支持扩充的并行关系数据库可以很好的解决以上数据仓库的技术,是数据仓库的核心,采用关系数据库实现的联机分析应用称为 ROLAP。

3.2.3 数据仓库的逻辑结构和物理结构

数据仓库是存储数据的一种组织形式,它从传统数据库中获得原始数据, 先按辅助决策的主题要求形成当前基本数据层,再按综合决策的要求形成综合数据层 (又可分为轻度综合层和高度综合层)。随着时间的推移,由时间控制机制将当前基本数据层转为历史数据层。可见数据仓库中逻辑结构数据由 3 层到 4 层数据组成,它们均由元数据组织而成。

数据仓库中数据的物理存储形式分为:基于关系数据库存储形式、多维数据库存储形式和虚拟存储形式。

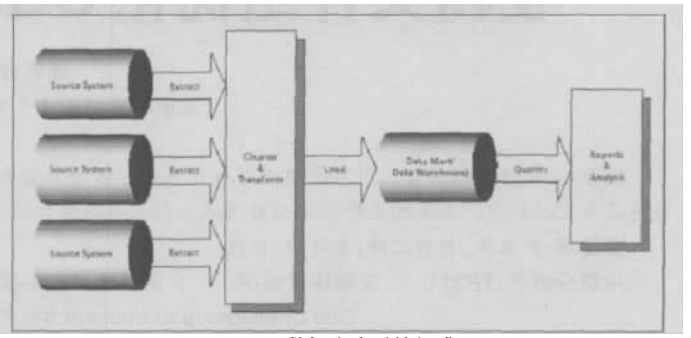

图 1 数据仓库系统组成

3.2.3.1 基于关系数据库的存储形式

基于关系数据库的存储形式就是将多维数据库的多维结构划分为两类表:一类是事实表,用来存储数据和维关键字;另一类是维表,即对每个维至少使用一个表来存放维的层次、成员类别等维的描述信息。维表和事实表通过主关键字和外关键字联系在一起,形成"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"。"星型模式"存在数据冗余、多维操作速度慢的缺点,但这种方式是主流方案,大多数数据仓库集成方案都采用这种形式。

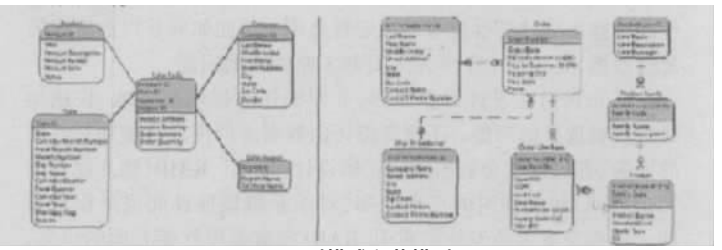

图 2 星型模式和花模型

3.2.3.2 多维数据库存储形式

多维数据库(MultiDimesional Database, MDDB)存储形式就是以多维的方式存储数据,以多维的方式来显示数据,即将数据存放在一个 n 维数组中。"维"是人们观察客观世界的角度,是一种高层次的类型划分。"维"一般包含着层次关系。多维数据在存储中将形成"超立方块(Hypercube)"的结构。当使用多维数据库作为数据仓库的基本数据存储形式时,减少了以维为基本框架的存储空间,针对多维数据组织的操作算法,极大地提高了多维分析操作的效率。

3.2.3.3 虚拟存储方式

虚拟存储方式是虚拟数据仓库的数据组织形式。它没有专门的数据仓库来存储数据,只是把指针存储于中心位置,而数据仍然在源数据库中, 只是根据用户的多维需求及形成的多维视图,临时在源数据库中找出所需要的数据,完成多维分析,数据源可以被实时地组合、传输和显示,而不必进行数据移动和复制,对于数据源也无须做任何改变。它让用户既能实时地看到历史数据,同时也能实时地看到当前数据。

3.2.3.4 几种存储形式的比较

多维数据库对多维概念表达清楚, 占用的存储空间较小,而且数据的综合速度高,但是多维数据库管理系统缺乏标准,而且多维数据库管理大规模数据库的能力不够强大。

基于关系数据库的存储形式,在灵活性和处理大规模数据的能力上完全可以满足数据仓库的需要。其不足在于数据库中存放

(下转第 104 页)

```
vunit M_edetect(M)
{
//检查错误检查能力
property pCheck1=always((EC&(^ED))- >next HE););
//EDA 奇偶校验
assert pCheck1;  //验证控制状态
property pCheck2=always(~(^)- >next HE);
assert pCheck2;  //验证数据通路
}
```

A. Check1:当 EC/ED 被驱动并且当通过 ED 被输入的数据是非法的时候，HE 在下一个循环中为真。EC 应该被限定在每个状态基和计数器范围内，但 ED 可以被他们共享。

B. Check2:当 I 是非法时 HE 在下一循环中为真。

(2)中间状态的合理性

第二个特性是中间状态的合理性检查，这个模块是检查当输入信号是完整的时候，中间状态是否也完整。

当没有错误数据输入并且输入是完整时，HE 不应该被断言。代码如下：

```
vunit M_integrity(M)
{
//完整性检查
property pIntegrityI=always(^);  //奇偶校验
assume pIntegrityI;  //假设输入完整
property pNoErrinjection=always(~EC);
assume pNoErrinjection;  //假设没有错误输入
property pIntegreityO=always(^O);//控制输出完整性
assert pIintegrityO;  //验证
}
```

(3)输出数据的完整性

第三个特性是输出数据的完整性。如果只要输入向量具有完整性那么输出向量就具有完整性，那么该特性就被设计。

当输入没有错误并且具有完整性时，那么输出也具有完整性。

4.3 系统的仿真验证

当模块的形式验证完成后，就进行系统的仿真验证，向系统随机输入测试向量，观测检查结果。

4.4 结果

实际证明该方法切实有效。不仅可以保证 100%逻辑功能的正确性，而且由于形式验证由电脑自动完成，可以节省很多时间，提高了工作的效率。

5 结束语

本文主要是解决像 SOC 这样超大规模设计的验证问题，提出了一种基于断言形式验证和仿真验证结合的混合验证方法的思想。核心思想是拆分系统，针对不同层次采用不同方法。将系统拆分为子模块，分别进行特性提取，并进行断言形式验证，保证模块正确。拆分后的模块级采用形式验证，可以避免状态爆炸的问题。然后仿真验证主要针对系统级，验证模块间的连接关系。这样测试向量就有针对性，而不是毫无目的编写 Testbench。最终可以达到 100%的覆盖率和正确性。

参考文献：

[1]PSL Language Reference Manual, version 1.01, 2003http://www.eda.org/vfv/docs/psl_lrm-1.01.pdf

[2]Tom Schubert. High Level Formal Verification of Next-Generation Microprocessors[J].InProceedings of DAC,2003,1-6.

[3]Yasushi Umezawa, Takeshi Shimizu A Formal Verification Methodology for Checking Data Integrity.

[4]Rolf Drechsler Towards Formal Verification on the System Level[M].Institute of Computer Science University of Bremen.

[5]韩伟华.基于断言的硬件设计功能验证技术[J].电子设计技术.

[6]admin 数字芯片设计的断言验证[J].中国集成电路杂志.

[7]韩俊刚,杜慧敏.数字硬件的形式化验证[M].北京:北京大学出版社,2001.

(上接第 33 页)

了大量的细节数据和相对较少的综合数据，需要以牺牲效率为代价动态地综合数据。

虚拟存储形式虽然较简单、花费少、使用灵活，但同时它只有当源数据库的数据组织比较规范、没有数据不完备及冗余，同时又比较接近多维数据模型时，虚拟数据仓库的多维语义层才容易定义，在实际中这种方式很难建立起有效的决策服务数据支持。

由于多维数据库管理系统及虚拟数据仓库技术的相对不成熟，关系数据库系统的广泛应用，目前在数据仓库的应用上基于关系数据库的存储形式仍占据着主流地位。

3.2.4 数据仓库数据的增量追加

定期向数据仓库追加数据也是十分重要的。数据仓库的数据是来自 OLTP 的数据库，一般可通过日志文件来确定需追加的新生数据。

3.2.5 数据仓库成本

通常情况下，在创建数据仓库时，实际收益是无法预测的，因为数据仓库的使用是用一种与其他信息处理不同的模式进行的，数据仓库是渐进式地建立的。第一次循环设计过程能很快完成，并且只需相对较少的费用。一旦数据仓库的第一部分已经建立并载入数据，分析员才能开始研究可能性，证明仓库开发费用的合理性。

3.2.6 数据仓库清理

数据并非只是注入数据仓库，它在数据仓库中也有自己的生命周期。到了一定时候，数据将从仓库中清除或上升到更高的综合级。数据清理或数据细节转化主要有以下几种方式：

(1)数据加入到失去原有细节的一个轮转综合文件中。

(2)数据从高性能的介质(如 DASD)转移到大容量介质上。

(3)数据从系统中实际清除。

(4)数据从体系结构的一个层次转到另一个层次，比如从操作型层次转到数据仓库层次。

因而，在数据仓库环境之中有种种数据清理或者转化的方式。数据的生命周期(包括清除或最终档案转移)应该是数据仓库设计过程中活跃的部分。

4 结束语

数据仓库是上世纪 90 年代发展起来的新技术，在我国的应用还刚刚起步，随着数据库技术的应用和发展，数据的累积越多，对历史数据的进行分析以提供决策依据的需求越大，数据仓库的应用需求也随之增大，构建高效的数据仓库增强企业的竞争力，同时对信息产业的发展有重要的作用。

参考文献：

[1]吴宏,陈奇,愈瑞钊.关于数据仓库若干问题的讨论[J].计算机科学,1999,2:39-43.

[2]陈波.基于数据仓库的决策支持系统的构建[J].电脑开发与应用,2002,9:23-27.

[3]周丽娟,邓颖,柳池.数据仓库技术和 OLAP 研究[J].佳木斯大学学报,2001,3:223-225.

[4]王珊.数据仓库技术和联机分析处理[M].北京:科学出版社,1995.

[5]李子木,莫倩,周兴铭.数据仓库的研究现状及未来访问[J].计算机科学,1998,4:57-58.