常用 OLAP 查询优化方法性能分析

张银玲,武 彤
(贵州大学 计算机科学与信息学院,贵州 贵阳 550025)

摘 要:OLAP(Online Analytical Processing)查询常常涉及到不同的维表和事实表,要得到查询结果通常需要进行多张表的连接操作。连接操作是一种非常耗时的操作,因此,如何提高 OLAP 查询效率成为数据仓库应用中的关键问题。文中对存储过程、索引技术、物化视图等几种常用的 OLAP 查询优化方法进行性能分析,针对特定应用通过反复实验比较得出物化视图的优越性。而就物化视图而言,其本身有优越性的同时也存在一些缺陷。因此,针对物化视图更新问题提出了几种更新方案。

关键词:OLAP;存储过程;索引技术;物化视图
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2014)01-0039-04
doi:10.3969/j.issn.1673-629X.2014.01.010

Performance Analysis of Several OLAP Query Optimization Methods

ZHANG Yin-ling, WU Tong
(College of Computer Science and Information, Guizhou University, Guiyang 550025, China)

Abstract: OLAP query often involves different dimension tables and fact tables, to get the query result usually requires more table join operations. The connection operation is very time-consuming, consequently how to improve the efficiency of OLAP queries becomes the key problem in data warehouse. Analyze several OLAP query optimization techniques like stored procedure, index technique, and materialized view, take the superiority of materialized view by doing experiments in application-specific system. And in terms of materialized views, it has the superiority, also has some defects. So, the several methods to solve the update problem of materialized view are proposed.

Key words: OLAP; stored procedure; index technique; materialized view

0 引 言

基于关系数据库的 OLAP(ROLAP)以关系数据库为核心,用关系二维表来存放数据。ROLAP 将多维数据库中的多维结构表划分为两类,一类是事实表,用来存储事实的度量值及各维的关键字以作为外键;另一类是维表,维表即用户分析问题的一个角度,简单说即一个方面。以事实表为中心,通过外键与若干维表联系在一起形成星型模型[1]。星型模型中,如果维度表是分层次的便组成雪花模型;如果事实表共享某些维度表便组成星座模型;而如果维度表分层次,同时事实表共享部分维表,则组成雪暴模型。无论是哪种模型,都是星型模型的扩展。星型模型是 OLAP 中被广泛采用的一种模型,该模型利于查询功能的实现,但在查询性能方面有很大的不足。用户每做一次查询,如果涉及到不同的维表和事实表,那么就要对不同的维表和事实表进行连接处理,而连接处理会花费很多额外的时间开销。文中基于一个实际的应用项目,该项目中的 OLAP 模型为星座模型,用户的查询请求大多需要连接 3 张或 3 张以上的维表和事实表。目前,实现 OLAP 查询的方式是基于视图的查询。这种查询方式只是简化了用户查询方式,在查询的效率上并没有作特殊处理。因此,寻找适合实际应用项目的 OLAP 查询优化方法非常必要。文中对存储过程、索引技术、物化视图等几种常用的 OLAP 查询优化方法进行性能分析,在该项目中通过反复实验对比得出物化视图的优越性。但是物化视图本身也存在许多缺陷,对此文中就物化视图更新问题提出几种物化视图更新方案。

收稿日期:2013-03-17 **修回日期**:2013-07-05 **网络出版时间**:2013-11-12
基金项目:贵州省工业攻关项目(黔科合 GY 字[2010]3061)
作者简介:张银玲(1988-),女,贵州铜仁人,硕士,CCF 会员,研究方向为数据库技术与应用系统;武 彤,硕导,教授,研究方向为数据库、数据仓库、数据挖掘技术。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20131112.1650.041.html

1 常用 OLAP 查询优化方法

1.1 存储过程

存储过程是由流控制和 SQL 语句书写,经编译和优化后存储在数据库服务器中[2],是一组能够完成特定功能的 SQL 语句集。它可以直接被调用,允许用户声明变量,可以接收和输出参数,返回执行存储过程的状态值,也可以嵌套调用。这些特性类似于编程技术中的类的某些特性(封装性、多态性和重构性)。存储过程可以重复使用从而减少开发者的工作量,同时也可指定特定的用户,增加安全性。在 OLAP 查询中,可以利用存储过程的以下两个优点来提高 OLAP 查询效率。

1)存储过程经过了编译解析和优化,可提高重复查询效率;

2)存储过程只需通过网络向服务器发出其名称和参数即可,而不是去传递和执行多关键字构成的 SQL 结构化语句符号串,这降低了网络通信量,提高了系统的响应速度。

1.2 索引技术

索引是一种树状结构,其中存储了关键字和指向包含关键字所在的记录的数据页的指针。当基于索引查找时,系统沿着索引的树状结构,根据索引中的关键字和指针,找到符合查询条件的记录。最后将全部查找到的符合查询语句条件的记录显示出来,避免了全表扫描。常用的索引方法有 B-树索引、位图索引和哈希索引,这些索引方法在数据仓库中已经得到了广泛的应用[3]。同时,新的数据仓库索引技术也在不断发展。主要包括 R-树索引、广义索引、位切片索引、标识技术、连接索引等等[4-5]。在 OLAP 查询中,使用索引不仅可以大大加快数据的检索速度,还可以加快 OLAP 查询时的连接速度。

1.3 物化视图

物化视图又叫实视图[6]或实体化视图[7],通过预先将那些费时的表连接操作或聚集操作进行预算并保存起来,是数据仓库中提高查询响应速度的一种高效技术。在 OLAP 查询优化中,可先按业务需要、相关主题或者统计分析等角度将数据抽取到物化视图中保存起来,然后让用户基于物化视图进行查询。这样做避免了从整个数据中寻找用户所需数据,这是一种典型的用空间换取时间的优化技术,它相当于部分远程数据的本地副本。物化视图对应用透明,增加和删除物化视图不会影响应用程序中 SQL 语句的正确性和有效性。

2 常用 OLAP 查询优化方法性能比较

2.1 OLAP 模型简介

文中基于实际的应用项目—电视机生产线决策分析系统。系统中以“电视机质量”为分析主题的 OLAP 模型如图 1 所示。

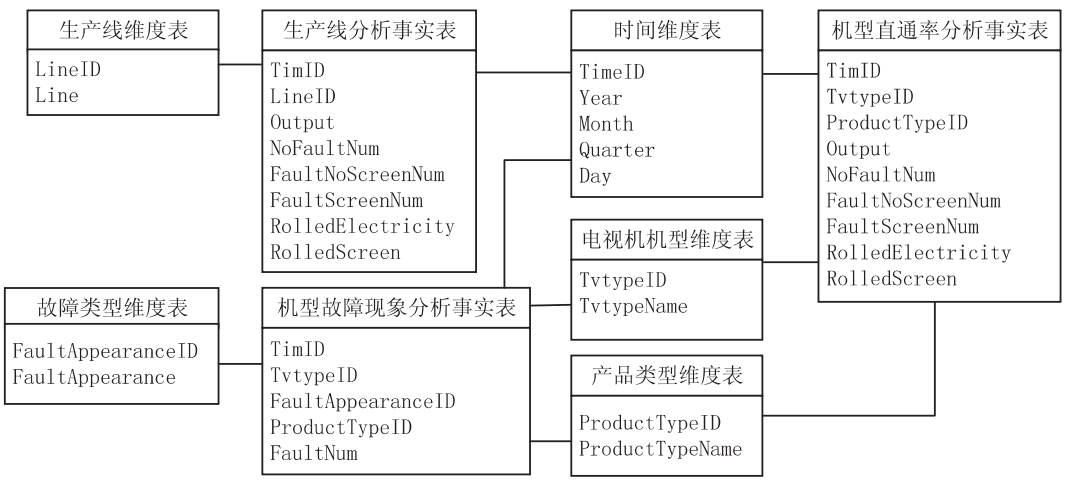

图 1 电视机质量分析主题 OLAP 模型

该模型中有 3 个事实表和 5 个维表,其中部分事实表共享同一个维表,组成星座模型。

· 生产线维度表:Dim_Line(LineID,Line);

· 故障类型维度表:Dim_FaultType(FaultAppearanceID,FaultAppearance);

· 时间维度表:Dim_Time(TimeID,Year,Month,Quarter,Day);

· 电视机机型维度表:Dim_TVtypeName(TvtypeID,TvtypeName);

· 产品类型维度表:Dim_Producttype(ProductTypeID,ProductTypeName);

· 生产线分析事实表:Productline_fact(TimID,LineID,Output,NoFaultNum,FaultNoScreenNum,FaultScreenNum,RolledElectricity,RolledScreen);

· 机型故障现象分析事实表:TvtypeName_product_FaultAppearance_fact(TimeID,TvtypeID,FaultAppear-

anceID,ProductTypeID,FaultNum);

· 机型直通率分析事实表:TvtypenameRolled_fact(TimeID,TvtypeID,ProductTypeID,Output,NoFaultNum,FaultNoScreenNum, FaultScreenNum, RolledElectricity,RolledScreen)。

2.2 实验内容

现在,选择不同的时间,不同类型,及不同机型来查询统计总产量(Output)、无故障数目(NoFaultNum)、非屏幕故障数目(FaultNoScreenNum)、屏幕故障数目(FaultScreenNum)、电性能直通率(RolledElectricity)、屏幕直通率(RolledScreen)等指标信息。要得到查询结果涉及到时间维度表、产品类型维度表、电视机机型维度表和机型直通率分析事实表 4 个表。

如查询 2008 年 1 月 1 日 至 2012 年 12 月 31 日 之间的平板产品和 LED2401 机型对应的技术指标(总产量、非屏幕故障数目、屏幕故障数目)。

直接多表查询 SQL(生成笛卡儿积)如下(元组数:148 803):

```
SELECT T.[Year],T.[Month],T.[Day],T.[Output],T.FaultNoScreenNum,T.FaultScreenNum
FROM (SELECTA.*,B.TVtypeName,C.*,D.*
FROM TVtypenameRolled_fact as A,DIM_TVtypeName as B,DIM_Time as C,DIM_Producttype as D
WHERE A.TimeID=C.TimeID AND A.TVtypeID=B.TVtypeID AND
A.ProductTypeID=D.ProductTypeID
)as T
WHERE T.[Year]<=2012 AND T.[Year]>=2008 AND T.[Month]IS NOT NULL AND T.[Quarter]IS NULL AND T.[DAY] IS NOT NULL AND T.TVtypeName='LED24K01'AND T.ProductTypeName='平板产品'
ORDER BY 1,2,3
```

将该 SQL 语句命名为 Q 语句。

运行 Q 语句 5 次的时间开销(单位:Microsecond)分别为:153,165,202,183,226;平均值为 185.8。

2.3 以存储过程进行查询优化

编写能够实现上述查询请求的存储过程。

```
CREATEPROCEDURE p_技术指标
AS
BEGIN
SELECTT.[Year],T.[Month],T.[Day],
T.[Output],T.FaultNoScreenNum,T.FaultScreenNum
FROM (
SELECTA.*,B.TVtypeName,C.*,D.*
FROM TVtypenameRolled_fact as A,
DIM_TVtypeName as B,
DIM_Time as C,
DIM_Producttype as D
WHERE A.TimeID=C.DIM_TimeID AND
A.TVtypeID=B.TVtypeID AND
A.ProductID=D.ProductTypeID
)asT
WHERET.[Year]<=2012 AND
T.[Year]>=2008 AND
T.[Month]IS NOT NULL AND
T.[Quarter]IS NULL AND
T.[DAY]IS NOT NULL AND
T.TVtypeName='LED24K01'AND
T.ProductTypeName='平板产品'
ORDERBY 1,2,3
END
GO
```

执行存储过程:EXEC[dbo].[p_技术指标]5 次时间开销(单位:Microsecond)分别为:210,130,129,133,165;平均值为:153.4。

2.4 以索引技术进行查询优化

分别在 DIM_Time 的 Year,Month,Day,Quarter 字段、DIM_TVtypeName 的 TVtypeName 字段、DIM_Producttype 的 ProductTypeName 字段上建立索引。

```
CREATE UNIQUE CLUSTERED INDEX Index_time
ON DIM_Time([Year],[Month],[Day],[Quarter])

CREATE UNIQUE CLUSTERED INDEX Index_TVtypeName
ON DIM_TVtypeName(TVtypeName)

CREATE UNIQUE CLUSTERED INDEX Index_Producttype
ON DIM_Producttype(ProductTypeName)
```

重新运行 Q 语句 5 次时间开销(单位:Microsecond)分别为:124,102,172,88,116;平均值为 120.4。

2.5 以物化视图进行查询优化

先建立一个能实现上述查询功能的物化视图 View_TVtypenameRolled_fact。然后基于物化视图进行相同的查询。

```
SELECT T.[Year],T.[Month],T.[Day],
       T.[Output],T.FaultNoScreenNum,T.FaultScreenNum
FROM   View_TVtypenameRolled_fact as T
WHERE T.[Year] <=2012 AND
       T.[Year] >=2008 AND
       T.[Month] IS NOT NULL AND
       T.[Quarter] IS NULL AND
       T.[DAY] IS NOT NULL AND
       T.TVtypeName = 'LED24K01' AND
       T.ProductTypeName = '平板产品'
ORDER BY 1,2,3
```

在物化视图上运行实现 Q 语句查询结果的查询 5 次时间开销(单位:Microsecond)分别为:112,99,101,106,99;平均值为:103.4。

2.6 性能对比分析

表 1 所示是以上三种方法的查询性能比较。从表 1 可以看出,同样的查询请求,使用不同的 OLAP 优化技术所用的开销不同,物化视图是其中最高效的方案。

表 1 性能分析对比

OLAP 查询技术	直接连接多表查询	存储过程	索引技术	物化视图
耗时/Microsecond	185.8	153.4	120.4	103.4

直接进行连接操作的查询方案是最原始最简单的操作方式,这种方式并没有对查询进行特殊的处理,因此当有多张表要进行连接时,将产生笛卡尔积;当数据量为无限多时,要获取到请求数据非常困难。存储过程的执行计划可以被缓存在内存中较长时间,减少了重新编译的时间,利用其缓存时间较长来降低短时间间隔内重复查询的时间开销,因此使用存储过程比直接运行 SQL 语句速度快。而 OLAP 分析常常用于管理者从整体宏观分析,其请求时间间隔往往比较长,因此存储过程也是不可行的。索引和物化视图一样是提高系统性能的最有效技术[7],但是如果能从较少的物化视图里抽取数据,性能明显高于直接从比物化视图数据量大的二维表中提取。数据仓库中若存储了大量的实视图,在处理查询时,系统可以选择处理一组代价小的实视图作为查询对象,而不用存取源关系表[8]。通过文中所做的实验,充分说明了物化视图比其他 OLAP 查询方案性能高。

3 物化视图的更新方案

通过物化视图来提高查询效率是典型的用空间换取时间的策略。这种方式提高了用户查询的时间,但是也存在很多不足之处,如占用存储空间,不能很好地适应用户查询要求发生变化以及物化视图选择和更新问题。针对文中的特定应用需求,发现查询要求变化不大,而一般企业采用的数据库服务器硬件配置都较高,所以存储空间也不是最关键的问题。而目前已经研究出基于聚类[9-10]、遗传[11-12]等大量的物化视图选择算法,因此最严峻的问题就是,当数据发生变动时,如何对物化视图进行更新。

目前,已经有许多物化视图更新方案[13-14],在主流的数据库管理系统中,主要存在两种物化视图的更新机制。一种是完全更新机制(Complete),在基本表数据发生更新时,删除物化视图中的所有数据而根据物化视图的定义重新生成物化视图,这种更新方式的缺点是当数据量很大时,更新速度极慢,性能损失严重;另一种是快速刷新机制(Fast),只将上一次对基本表处理以后的新查询数据增量更新到物化视图中,但这种方式有一定的条件限制,并非所有的物化视图都能快速刷新。

而针对特定应用系统,文中提出以下几种方案旨在能够较好地解决物化视图的更新问题。

(1)根据用户的查询频率,动态更新物化视图。如果能准确收集到用户的查询信息,那么这种方案是一种最快最好的更新方式。这种方式属于快速刷新机制,但是其数据量会远远少于快速刷新机制,根据用户查询频率反应出的查询信息将过滤掉那些多余的冗余数据。

(2)定义一个更新间隔时间,建立存放时间间隔内的中间视图,定期将中间视图的数据更新到物化视图中。这种方式解决了快速刷新机制的局限性,适合任何物化视图。

(3)编写专门的处理程序部署于数据库服务器,当有新数据更新时,触发器自动激活该处理程序并将 OLAP 中的数据按照用户查询需求导入到物化视图中,以实现物化视图数据的更新。这种方式用程序来代替人工刷新,解决了何时更新物化视图数据的问题。

4 结束语

在众多的 OLAP 优化技术当中,如何选择合适的优化技术来提高信息系统的查询效率是一个很关键的问题。不同的 OLAP 查询优化技术有其适用情况,因此,在选择时应结合系统的特点来选择不同的优化技术。当为高基数数据时应选择 B 树索引,低基数时则适合用位图索引。综合来看,如果不考虑系统存储空间,物化视图是提高检索性能较好的解决方案。使用物化视图必须要考虑其数据更新问题,文中提出了几种解决物化视图更新问题的方案,每种方案都在一定的假设前提下,因此在实际的设计时可能存在某些不适用性。将在今后工作中更加深入研究物化视图更新方案,同时通过实验去验证其适用性。

参考文献:

[1] Inmon W H. Building the data warehouse[M]. [s. l.]: Wiley India Pvt, Limited, 2005: 126-133.

[2] 王 珊, 萨师煊. 数据库系统概论[M]. 北京: 高等教育出版社, 2009: 247-248.

[3] Agost L. 数据仓库技术指南[M]. 北京: 人民邮电出版社, 2001: 76-78.

[4] 周丽娟, 刘大昕, 柳 池, 等. 数据仓库技术在 CIMS 环境中的应用研究[J]. 计算机应用与软件, 2003, 20(3): 62-64.

[5] 谢立宏, 贺贵明, 任小荣. OLAP 数据的索引[J]. 微型机与应用, 2002, 21(11): 11-14.

[6] Theodiratosd B M. A general framework for the view selection

(下转第 46 页)

结构,对于多瓶颈链路拓扑结构却少有涉及。在单瓶颈链路拓扑结构条件下研究各种算法理论上是可行的,但是现实的网络结构却是非常复杂的,很多因素的影响是单瓶颈链路所无法模拟的,而这些因素很可能是影响实际网络性能的重要因素,多瓶颈链路拓扑是一种更接近现实网络模型的结构,只有在多瓶颈链路条件下对 TCP 网络拥塞控制算法进行研究,才能更好地揭露算法在实际网络中可能涉及到的难题,因而对多瓶颈链路网络进行深入研究对于实现网络的拥塞控制具有重要意义[12]。

3.4 基于控制理论的算法研究

控制理论是一门相当成熟的理论,有非常多的方法可以借鉴到拥塞控制中来。近年来国内外的很多学者进行了一些尝试性工作,利用控制理论的方法来解决互联网中的拥塞控制问题。但是由于 Internet 本身是一个复杂非线性结构,使对网络稳定性和动态性能的分析的研究更加困难,因而这方面的研究还不够成熟,有待继续研究。因此,如何有效地将控制理论的思想特别是智能控制方法运用于网络拥塞控制中,将是未来研究的一个难点问题,也是一个热点问题[13]。

3.5 基于智能优化算法的拥塞控制研究

智能优化算法又称为现代启发式算法,是一种具有全局优化性能、通用性强且适合于并行处理的算法。这种算法一般具有严密的理论依据,而不是单纯凭借专家经验,理论上可以在一定的时间内找到最优解或近似最优解。近年来将智能优化算法应用于拥塞控制也成为了一个热门的研究方向,它主要用来解决那些传统方法无法解决的拥塞控制问题。例如:遗传粒子群优化算法、蚁群优化算法、多 Agent 算法在拥塞控制中的研究已经取得了初步的进展。

4 结束语

文中在介绍网络拥塞与拥塞控制的基础上,重点介绍了 TCP 网络拥塞控制的原理、控制机制、控制算法。列举出了现有算法,并对其进行分析,对这些算法的优缺点进行了比较。总结了 TCP 拥塞控制目前的研究成果,指明了未来 TCP 控制研究热点的发展方向,为 TCP 控制接下来的研究奠定了基础。

参考文献:

[1] Floyd S, Jacobson V. Random early detection gateways for congestion avoidance[J]. IEEE/ACM transactions on networking, 1993, 1(4): 397-413.

[2] Wang Y J, Schinkel M, Schmitt-Hartmann T. PID and PID-like controller design by pole assignment within D-stable regions[J]. Asian journal of control, 2002, 4(4): 41-52.

[3] 孔金生,赵长伟,万百五. 网络拥塞的智能化适应控制方法[J]. 系统工程与电子技术,2005,27(7):1301-1303.

[4] 王满喜,胡向晖,马刈非. 混合式的网络拥塞控制算法[J]. 电子科技大学学报,2007,36(3):642-645.

[5] 刘 翔. 基于智能方法的网络拥塞控制技术的研究[D]. 天津:天津工业大学,2007.

[6] 王秀利. 网络拥塞控制及拒绝服务攻击防范[M]. 北京:北京邮电大学出版社,2009.

[7] 陈尚兵,王 彬,钱积新. TCP 拥塞控制综述[J]. 计算机科学,2002,29(5):32-35.

[8] 武航星,慕德俊,潘文平,等. 网络拥塞控制算法综述[J]. 计算机科学,2007,34(2):51-56.

[9] 刘 俊,谢 华. 一种改进的 TCP 拥塞控制算法[J]. 计算机工程,2011,37(13):95-97.

[10] 李 涛. 改进 TCP 拥塞控制算法的仿真及应用研究[J]. 计算机仿真,2011,28(6):181-184.

[11] 桂晓琳,许向阳. 基于 Elman 神经网络的网络流量预测[C]//2005 年全国自动化新技术学术交流会会议论文集. 出版地不详:出版者不详,2005:288-291.

[12] 徐小卜. 一类具有三条瓶颈链路的网络系统稳定性分析[J]. 电脑与电信,2011(11):46-48.

[13] 汪小帆,孙金生,王执铨. 控制理论在 Internet 拥塞控制中的应用[J]. 控制与决策,2002,17(2):129-134.

(上接第 42 页)

problem for data warehouse design and evolution[C]//Proceedings of the ACM third international workshop on data warehousing and OLAP. Mclean, VA, USA: [s. n.], 2000.

[7] Golfarelli M, Rizzi S. 数据仓库设计:现代原理与方法[M]. 北京:清华大学出版社,2010:229-233.

[8] 周丽娟,柳 池,刘大昕. 在数据仓库中使用实视图优化查询[J]. 计算机工程与应用,2004,40(16):181-183.

[9] 吕 晓,陈 耿,朱玉全. 基于聚类的动态物化视图选择研究[J]. 计算机工程与设计,2009,30(15):3638-3640.

[10] 梁 银. 基于聚类方法的空间度量物化选择算法[J]. 计算机工程,2011,37(8):58-60.

[11] Lawrence M. Multiobjective genetic algorithms for materialized view selection in OLAP data warehouses[C]//Proc of GECCO'06. Seattle, Washington, USA: [s. n.], 2006.

[12] 王宜贵. 基于遗传算法的物化视图优化方法[J]. 计算机与现代化,2011(8):23-25.

[13] 武 彤,赵 雪,赵 洵. 动态更新实物化视图以提高 OLAP 查询效率[J]. 计算机科学,2012,39(B06):315-317.

[14] Colby L S, Griffin T, Libkin L. Algorithms for deferred view maintenance[C]//Proc of SIGMOD'96. Montreal, Canada: [s. n.], 1996.

常用OLAP查询优化方法性能分析

作者： 张银玲， 武彤， ZHANG Yin-ling， WU Tong
作者单位： 贵州大学 计算机科学与信息学院,贵州 贵阳,550025
刊名： 计算机技术与发展
ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(1)