

云计算环境中多维数据集的查询与分布策略研究

潘惠勇¹, 高丽平¹, 刘 扬²

(1 中原工学院 计算机学院, 河南 郑州 450007; 2 河南工业大学 科研处, 河南 郑州 450001)

摘 要: 针对多维数据集的特点, 首先定义了查询等价集等概念, 并结合云计算环境提出了多维数据集的查询算法. 根据算法过程中各类因素的影响, 提出了一种云计算环境中多维数据集的数据分布策略. 理论分析与实验结果均表明, 将本文提出查询与分布策略应用于云计算环境中的多维数据集, 能有效的优化数据分布, 提高数据查询的效率.

关键词: 云计算; 多维数据集; 数据分布

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2012)08-0024-04

Research on Query and Data Distribution Strategy of Data Cube in Cloud Computing Environment

PAN Hui-yong¹, GAO Li-ping¹, LIU Yang²

(1 School of Computer Science, Zhongyuan University of Technology, Zhengzhou 450007, China;

2 Department of Scientific Research, Henan University of Technology, Zhengzhou 450001, China)

Abstract: This paper define the query equivalent set by analyzing the character of the data cube firstly and provide a query algorithm of data cube in cloud computing environment. A data distribution strategy of data cube in cloud computing environment is illustrated in the query procedure of inspecting the effect of various factors. Theoretical analysis and experimental results show that the proposed distribution strategy and the query applied to data cube in cloud computing environments can effectively optimize data distribution and improve the query efficiency.

Key words: Cloud Computing, Data Cube, Data Distribution

1 引言

云计算是一种随着计算机技术的发展和應用而形成的互聯网络, 它是一种可以动态伸缩, 并且经常可以提供虚拟资源作为服务以供使用的互联网. 文献[1]指出了云计算环境与网格等技术所区分开来的重要特点: 云计算资源的虚拟化, 即云计算环境中的资源可以根据用户的需求自行配置.

从数据的角度来看, 云计算环境与网格环境有相似之处^[2], 但云计算对于数据管理的目标是实现数据的优化使用, 因此需要有一个较好的数据管理与分布策略^[3], 能合理调度云系统中的资源来及时响应用户的需求, 并提高数据的安全性^[4]; 文献[5]

提出了一种云计算环境中基于查询的关系数据库分布策略. 多维数据集作为一种数据的组织形式^[6], 其在云计算环境中的查询与分布策略都与关系数据库有着很大的不同, 因此需要结合其本身的特性与云计算环境来进行综合研究.

2 基本概念

多维数据集通过其度量值和维度定义. 其中维度用来描述度量属性, 度量值一般为数值类型, 用于进行分析处理.

一个多维数据集由基表元组和若干个聚集元组构成. 为便于说明, 考虑一个描述产品销售量的多维数据集, 其包括三个维度: 时间、地区、产品; 一个度

收稿日期: 2012-01-15; 修回日期: 2012-03-20

基金项目: 国家自然科学基金项目(61003052)

量:销售量.其中时间维含有三个层次(年、季度、月),地区维含有两个层次(地区、省),产品维为单层次维.

定义1 多维数据集中的基表.多维数据集中的基表即为事实表,其各个字段根据语义区分为维(层次)及度量字段.

对于产品销售量的多维数据集,其基表如表1所示.其描述了各个时间点上各个地区的产品销售情况.

定义2 多维数据集中的聚集表.聚集表是基表按照层次,在各个维度上进行聚集,度量进行相应计算后,所得到的表.

表2即表1基表所对应的聚集表,其聚集函数为SUM.其中“*”表示“所有值”.

表1 多维数据集中的基表

年	季度	月	地区	省	产品	销售额
2011	1	1	华中	河南	A	100
2011	1	1	华中	湖南	A	50
2011	1	3	华中	河南	B	50
2011	1	1	华东	上海	A	60
2011	2	4	华中	河南	A	50
.....						

表2 多维数据集中的聚集表

年	季度	月	地区	省	产品	销售额
2011	1	*	华中	河南	A	100
2011	*	*	华中	河南	A	150
*	*	*	华中	河南	A	150
2011	1	1	华中	*	A	150
2011	1	1	*	*	A	200
2011	1	1	华中	河南	*	100
2011	1	*	华中	*	A	210
.....						
*	*	*	*	*	*	*

由上面的定义可知,多维数据集中的聚集元组个数与基表元组个数存在指数关系,一般都不可能计算出所有的聚集元组并加以存储,因此现有的各类多维数据集存储方案都是只保存部分的聚集元组,如果需要查询未预先计算保存的聚集元组,则根据一定的方式加以计算并返回结果.

定义3 单元格.多维数据集中的基表元组与聚集元组统称为此多维数据集中的单元格.

定义4 单元格的聚集距离.对于两个单元格 C_a, C_b ,如果:

① 两个单元格相同,则定义这两个单元格的聚集距离为0;

② 如果在某一个维上 D_i 的层次值或者在 C_a 上为“*”,或者相等,而在其他维上的值都相等,则定义 C_b 到 C_a 的聚集距离为 D_i 上“*”相差的个数;

③ 其他情况,则定义两个单元格之间的聚集距离为无穷大.

聚集距离描述了两个单元格是否存在聚集关系,并给出了这种聚集关系的一种度量.

定义5 单元格的基表元组集.如果一个单元格是由若干个基表元组聚集而来,则称这些基表元组的集合为这个聚集元组的基表元组集.并特别规定:对于一个基表元组,其对应的基表元组集为其本身.

定义6 单元格的查询等价集.对于一个单元格 C (基表元组集为 S),如果存在一个单元格集 $\{C_1, C_2, \dots, C_n\}$,其对应的基表元组集分别为 $\{S_1, S_2, \dots, S_n\}$,任取 S_i, S_j ,都满足, $S_i \cap S_j = \emptyset$,同时有 $\bigcup_{i=1}^n S_i = S$,则称这个单元格集为此单元格 C 的查询等价集,记为 $C_{=}$.

一个单元格可能存在多个查询等价集.在计算某个单元格的度量值时,如果其对应的某个查询等价集中的单元格已经计算,则可以通过对查询等价集的计算汇总而得到结果,避免从基表元组开始计算,从而有效的提高计算效率.

3 查询与分布算法

3.1 云计算环境中的多维数据集查询算法

传统的多维数据集存储一般采用C/S或B/S架构——服务器端存储维护多维数据集的单元格,客户端提交查询,服务器端对查询进行解析、匹配与计算汇总,生成查询结果后返回给客户端.而在云计算环境中,由于多维数据集的单元格分配在不同的数据节点上,因此传统的实现框架与查询算法都不适合,同时,云计算平台强调资源与数据的虚拟化,多维数据集的单元格有可能在不同的节点中进行配置与迁移,因此传统意义下的并行查询算法也不能直接应用于云计算平台中.针对云计算环境的特点,

本文设计了云计算环境下的多维数据集查询与分布算法。

算法 1 云环境下的多维数据集查询算法 ServerQuery(Q)

输入: Q, 为用户提交到服务器的多维数据集查询, 即一个不包含度量值的单元格

输出: Q 所对应的度量值

Begin

将查询 Q 分发到各数据存储节点(Node), 并设置集合 S_Q 为空;

各个节点查找与查询 Q 聚集距离为 0 或 1 的那些单元格, 并返回至服务器;

收集各个节点返回的结果, 并加入至 S_Q 中;

if ($Q \in S_Q$) then 返回结果; // 查询 Q 已经预计算并存储, 可以直接得到结果

else if ($Q \in S_Q$) then 计算汇总 Q_{sum} , 并返回结果; // 查询 Q 虽然并没有预计算并存储, 则可以通过计算其对应的某个查询等价类得到结果;

else // 对于那些等价查询集的子集, 生成与对应等价查询集的补集, 并选择其中元素最少的那个

{
选择某一个查询等价集的子集 CQ_{sum} , 其对应的补集 $\overline{CQ_{\text{sum}}}$ 所包含的元素最少;

计算汇总 CQ_{sum} 中每个单元格度量值, 结果记为 Q_i ;

对 $\overline{CQ_{\text{sum}}}$ 中的每个单元格 q_i , 计算汇总 ServerQuery(q_i), 并与 Q_i 进行汇总, 返回所得结果;

}

End

算法的思路就是将查询分发给各个数据存储节点, 如果查询的结果已经预先计算并存储在了某个数据节点上, 则可以直接给出结果; 如果一个查询本身没有被计算出来, 但可以找到其对应的一个查询等价集, 则可以通过计算其查询等价集得到查询结果, 否则就必须递归调用算法, 对某个查询等价集的缺失部分进行查询, 并将查询结果与已经找到的结果进行组合, 计算汇总后得到最终的结果。

3.2 云计算环境中多维数据集的数据分布策略

由算法 3.1 所描述的查询过程可知, 在云计算平台中, 对于一条多维数据集的查询, 其主要时间耗费为:

- (1) 查询在云计算平台中的网络传输时间 t_{net} ;
- (2) 在数据节点中查找与查询聚集距离为 0 或 1 的单元格, 记为 t_{node} ;
- (3) 服务器上对节点传回来的查询结果进行整

理与计算, 记为 t_s ;

(4) 对查询等价集缺失部分查询的处理时间 t_m .

因此, 总的查询时间可以记为 $t = t_{\text{net}} + \max(t_{\text{node}}) + t_s + t_m$. 式中, t_{net}, t_s 可以视为一个常数, 而对 t_m 的优化, 则在算法 3.1 中通过每次选择元素最少缺失集的方式进行; $\max(t_{\text{node}})$ 表示各个节点中最大的查找时间, 因此需要通过对用户提交的查询执行过程进行分析, 根据其中的信息来对各个数据节点的单元格进行均衡分布, 以减少 $\max(t_{\text{node}})$. 为了更好的描述同一个节点上单元格的关系, 定义单元格之间的相关度如下:

定义 7 单元格的相关度. 设 C_1, C_2 为两个单元格, list_{sum} 为当前所有查询所对应的查询等价集的集合, $\text{count}(C_1, C_2)$ 为 C_1, C_2 出现在同一个查询等价集中的次数, $\text{count}(C_1), \text{count}(C_2)$ 分别表示其在查询等价集集合中出现的次数, 则这两个单元格的相关度为 定义为 $\frac{\text{count}(C_1, C_2)}{\text{count}(C_1) + \text{count}(C_2)}$, 记为 $R(C_1, C_2)$.

其算法如下:

算法 2 基于查询频率的多维数据集单元格均衡分布算法

输入: 多个查询与其对应的查询等价集

输出: 各数据节点上单元格的分布调整方案

Begin

统计输入的查询等价集, 计算各个分数据节点的访问频率;

找出访问频度最高的节点 node_i ; // 访问频度高, 说明其访问量大, 需要将此节点上的单元格分布到其他数据节点中;

选择 node_i 中在所有等价查询集中出现次数最多的单元格 C , 并计算此节点与这个节点中其他单元格 C_i 的相关度 $R(C, C_i)$, 并进行排序, 选取一定比例的单元格, 按比例迁移到其他节点上去;

End

4 实验与分析

为了对文中的算法与分布策略进行实验分析, 搭建了一个云计算平台, 使用一台 PC 作为主服务器, 对多维数据集进行查询与分布控制. 同时使用 4 台 PC 作为数据存储节点, 配置均为 2.4GHz CPU, 2GB 内存, 操作系统为 Ubuntu 11.04. 实验用的多

维数据集采用某个大型跨国企业的销售数据.在初始阶段,预先计算了 10^4 个单元格,并在这些结点上进行随机分布.

实验一:对均衡分布算法进行了分析,其结果如图1.可以看出,随着每一次的重新分布,每个节点上的访问次数都趋向于平均,而且所参照的查询数越多,所分布的次数越多,均衡化的效果越明显.

实验二:对分布策略的优化作用进行了对比,每次随机生成 1000 个查询,记录平均查询时间,然后根据这 1000 个查询进行单元格分布,并与前次查询时间进行对比,其结果如图2所示.由图2可以看出,在每次分布后,响应速度都有了提升,随着参数查询数的提高与分布次数的增加,所耗费的平均时间与理论上的最优响应时间相差越来越小.

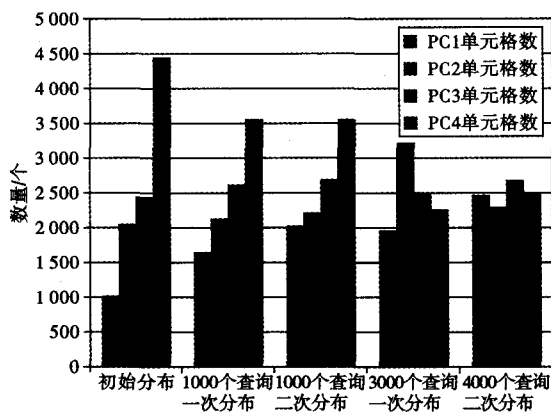


图1 均衡分布算法实验

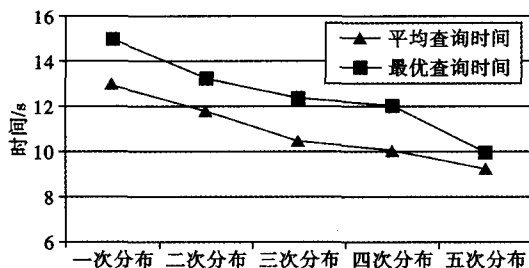


图2 分布策略优化实验

5 结束语

在云计算环境中,多维数据集的存储与分布方式与传统的关系型数据库截然不同,查询方式也需要进行调整与优化.本文结合多维数据集与云计算环境的特点,提出了云计算环境下的查询与分布策略,并进行了实验,取得了较好的优化效果.在后续的研究中,将针对云计算环境下的多维数据集的安全性,备份与冗余存储等方面进行研究.

参考文献:

- [1] Luis M Vaquero, Luis Rodero-Merino, Juan Caceres, et al. A break in the clouds: towards a cloud definition [J]. ACM SIGCOMM Computer Communication Renew, 2008, 38(1): 50-55.
- [2] Foster I, Zhao Y, Raicu I, et al. Cloud compming and grid computing 360-degree compared[J]. Grid Computing Environments Workshop, 2008: 1-10.
- [3] 宁康, 时成阁. 网络存储导论[M]. 北京: 清华大学出版社, 2007.
- [4] Amiri K, Gibson G A, Golding R. Highly concurrent shared storage distributed computing systems[C] // Proceedings of 20th International Conference. Washington DC, USA. 2000: 298-307.
- [5] 文明波, 丁治明. 适用于云计算的面向查询数据库数据分布策略[J]. 计算机科学, 2010(9): 168-172.
- [6] Chaudhuri S, Dayal U. An overview of data warehousing and OLAP technology[J]. ACM SIGMOD Record, 1997, 26(1): 65-74.

作者简介:

潘惠勇 男, (1977—), 硕士, 讲师. 研究方向为复杂网络与智能计算、云计算.

高丽平 女, (1976—), 硕士, 副教授. 研究方向为分布式计算.

刘扬 女, (1978—), 博士, 副教授. 研究方向为云计算、分布式计算、网格计算.