学校编码：10384　　分类号__密级__
学号：23020061152445　　UDC___

厦门大学

硕 士 学 位 论 文

数据仓库中物化视图的选择与调整

Selection and Adjustion of Materialized View in Data Warehouse

王金水

指导教师姓名：张东站副教授
专 业 名 称：计算机软件与理论
论文提交日期：
论文答辩时间：
学位授予日期：

答辩委员会主席：__

评阅人：__

2009 年 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范(试行)》。

另外，该学位论文为(　　)课题(组)的研究成果，获得(　　)课题(组)经费或实验室的资助，在(　　)实验室完成。(请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。)

声明人(签名)：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文(包括纸质版和电子版)，允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

(　　)1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

(　　)2.不保密，适用上述授权。

(请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。)

声明人(签名)：

年　　月　　日

摘要

数据仓库是一个面向主题的、集成的数据集合，用来支持管理人员的决策。它维护着海量的数据并且支持形式复杂的查询，如何高效的管理如此之多的数据并提供高效的查询是数据仓库面临的其中一个难题，而物化视图是解决这个问题的重要手段，但是它需要额外的系统空间来存储，并且需要牺牲系统的代价来维护，因此，物化视图的选择是一个重要的研究课题。传统物化视图的静态选择算法是基于查询分布概率已经由用户提供，或这些查询在综合数据上是均匀分布的前提下。但实际应用中，用户查询均匀分布或用户提供查询概率的假设常常不能成立，因此提出一种既有静态选择能力，又有动态调整能力的视图选择算法就显得相当有实际与研究意义。

本文从静态和动态两方面深入研究物化视图的选择问题，基于 SPJ（Select-project-join）视图假设的数据仓库模型，以 MVPP 为搜索空间，综合考虑存储空间、视图维护及查询性能，提出了一种新的物化视图选择和调整算法--VSAA(Views Selection and Adjustment Algorithm)。从数学模型和关键参数两个方面研究了 VSAA 的理论模型，针对 VSAA 实时性差的缺陷提出了 DVMV(Dynamic Virtual Materialized Views)算法作为 VSAA 算法的补充，以理论分析为指导实现了 VSAA 的算法并和各种经典算法进行了对比分析，利用 TPC-H 基准数据生成 1G 数据，并导入到 Oracle 数据库中作为实验数据，通过理论及实验证明了 VSAA 算法的有效性和优越性。

关键字：数据仓库；物化视图；静态选择；动态调整

Abstract

A data warehouse is, by definition, a subject-oriented, integrated collection of data which support managers to make correct decision. It enables complex queries on data warehouses, which contains large amounts of data. However, How to manage so much data efficiently and provide high efficient query is one of the problems faced by data warehouse. And materialized view is a important method to solve that, but it needs extra system space to store and also needs to be maintained at the price of system immolation. Therefore, The selection of materialized view is an important research subject. The static selection algorithm of traditional materialized view is based on the assumption that probability distribution of queries has been provided by the user, or these queries in the integrated data are uniformly distributed. But these assumptions always can not be realized in practical use. So a view selection algorithm with the capacity of both static selection and dynamical adjustment is quite practical and researchable.

This paper lucubrate the selection of materialized view from static and dynamic aspects and presents a new materialized view selection and adjustment algorithm --VSAA (Views Selection and Adjustment Algorithm) which based on the SPJ（Select-project-join）view assumption data warehouse model .This algorithm take MVPP(multi-view processing plan) as the search space and integrately consider the factors of storage space, view maintenance and query capability. After having researched the VSAA theoretical mode from mathematic model and key parameter aspects, this paper put forward DVMV (Dynamic Virtual Materialized Views) algorithm as a supplementary of VSAA algorithm specific to the real-time defect of VSAA. And take the theoretical analysis as guidance to realize the VSAA algorithm as well as contrastive analysis to other classics algorithms. Finally, I use tpc-h benchmark data to generate 1G data into Oracle database as the experimental data, and proved the effectiveness and superiority of the algorithm VSAA through theories and experiments.

Key Words：Data Warehouse; Materialized view; Static selection; Dynamic adjustment

目　　录

第一章 绪论 1

1.1 引言 1

1.2 联机分析处理概述 1

1.2.1 OLAP 的功能特征 3

1.2.2 OLAP 的实现 3

1.3 物化视图概述 5

1.3.1 物化视图的概念 5

1.3.2 物化视图的主要管理任务 6

1.4 国内外研究现状 11

1.4.1 物化视图静态选择算法 11

1.4.2 物化视图的动态调整 13

1.5 存在的问题 13

1.5.1 物化视图选择的负面因素 13

1.5.2 静态物化视图选择的缺陷 14

1.6 本文的工作 15

1.7 本文的组织结构 15

第二章 经典物化视图的选择与调整算法 17

2.1 多维物化视图的计算模型 17

2.1.1 多维物化视图的尺寸计算 18

2.1.2 多维物化视图的代价计算 18

2.1.3 多维物化视图的收益计算 20

2.2 静态选择算法 22

2.2.1 Greedy 算法 22

2.2.2 YKL 算法 23

2.2.3 IMDVSA 算法 24

第三章 基于 MVPP 的物化视图选择算法 VSAA 的理论基础 .27

3.1 VSAA 算法的数学模型 27

3.1.1 VSAA 的维护策略 27

3.1.2 VSAA 的调整策略.....28
3.1.3 VSAA 的视图表示.....28
3.1.4 VSAA 代价计算模型.....30
3.2 VSAA 关键参数.....32
3.2.1 物化视图的初始空间 SPACE.....32
3.2.2 查询访问集合 Q.....34
3.2.3 未命中查询率.....38
3.2.4 视图收益阈值.....39
第四章 基于 MVPP 的物化视图选择算法 VSAA 的算法实现 41
4.1 VSAA 算法实现.....41
4.1.1 VSAA 算法描述.....41
4.1.2 VSAA 算法进一步说明.....44
4.2 VSAA 算法分析.....45
4.2.1 VSAA 算法的理论分析.....45
4.2.2 VSAA 实验分析.....47
4.3 实验结论.....53
第五章 结束语.....55
5.1 总结.....55
5.2 下一步要做的事.....55
参考文献.....57
附录 数据仓库的数据填充.....63
使用 tpc-h dbgen 产生数据.....63
将产生的数据导入 oracel 数据库:.....65
攻读硕士学位期间发表的论文.....67
致谢.....69

Contents

Chapter 1 Introduction 1

1.1 Foreword 1

1.2 Online Analytical Processing overview 1

1.2.1 The features of OLAP 3

1.2.2 The realization OLAP 3

1.3 Materialized views overview 5

1.3.1 The concept of materialized views 5

1.3.2 The main management tasks of materialized views 6

1.4 Researches at home and abroad 11

1.4.1 The static selection algorithm of materialized view 11

1.4.2 Dynamic adjustment of materialized views 13

1.5 Problems 13

1.5.1 The negative factors of materalized view selection 13

1.5.2 The defects of static selection of materialized view 14

1.6 The task of this paper 15

1.7 The organizational structure of this paper 15

Chapter 2 Classic Materialized View Selection and Adjustment Algorithm 17

2.1 Muti-dimensional calculation model of materialized view 17

2.1.1 The size calculation of multi-dimensional materialized view 18

2.1.2 The cost calculation of multi-dimensional materialized view 18

2.1.3 The profit calculation of multi-dimensional materialized view 20

2.2 The static slection algorithm 22

2.2.1 Greedy algorithm 22

2.2.2 YKL algorithm 23

2.2.3 IMDVSA algorithm 24

Chapter 3 The Theoretical Basis of MVPP-based Selection Algorithm VSAA of Materialiazed Views 27

3.1 The mathematical model of VSAA algorithm 27

3.1.1 The maintenance strategy of VSAA 27

3.1.2 The adjustment strategy of VSAA 28

3.1.3 The expression of VSAA 28

3.1.4 The cost calculation model of VSAA 30
3.2 The key parameters of VSAA 32
3.2.1 The initial space of materialized view 32
3.2.2 The collection of the query 34
3.2.3 The rate of the missing hit query 38
3.2.4 The threshold of the view profit 39
Chapter 4 The Realization of VSAA Algorithm Based on MVPP 27
4.1 VSAA algorithm 41
4.1.1 The description of VSAA algorithm 41
4.1.2 The further illustration of VSAA algorithm 44
4.2 The analysis of VSAA algorithm 45
4.2.1 The theoretical analysis of VSAA algorithm 45
4.2.2 The experimental analysis of VSAA algorithm 47
4.3 Experimental Conclusions 53
Chapter 5 Epilogue 55
5.1 Conclusion 55
5.2 Next work 55
References 57
Appendix Data filling of data warehouse 63
Generate data with tpc-h dbgen 63
Export the generated data to oracel data warehouse 65
Publictiion 67
Acknowledgement 69

第一章 绪论

1.1 引言

在过去的三十多年里，数据库系统作为数据管理手段主要用于事物处理。这些数据库已经存放了大量的日常业务数据。尤其到了二十世纪末，信息技术发展迅速，数据量加倍的时间越来越短。从大量的数据中得到有价值的，支持决策制定的信息是信息系统用户的新需求。因此，数据仓库技术应运而生。

经过近几年的发展，数据仓库技术已经日臻完善，大多数大型数据库供应商都已在其产品中集成了数据仓库的功能。例如 Oracle、IBM DB2、Sybase、SQL Server 等都已经向用户提供了比较成熟的数据仓库产品，并且在实际应用中取得了显著的效果。数据仓库中数据的大量积累、用户数量的增加、查询的复杂化，使得数据仓库必须满足更高的性能要求。虽然数据仓库没有 OLTP 那么苛刻的响应时间要求（通常少于 3 秒），然而如果响应时间过长用户是不能接受的。

物化视图技术是数据仓库系统中提高性能的关键性技术之一，它是一种将视图所对应数据加以实际物理存储的技术。其目的是通过预计算来加快数据仓库系统对用户查询的响应速度。然而视图的物理化既需要占用可观的磁盘空间，又需要耗费大量的系统资源以对其进行维护。所以如何选择一组合适的视图集合加以物理化，使系统能够利用有限的资源最大限度的提高数据仓库系统对用户查询的响应速度，是一个极为重要的问题。本文主要研究数据仓库中的物化视图选择问题，此课题具有一定的理论意义和实用价值。

1.2 联机分析处理概述

数据仓库把来自多个数据源的数据集成起来，形成一个可靠的、一致的和不断更新的历史信息集合。数据仓库是进行分析决策的基础，但还必须有强有力的工具进行分析和辅助决策。基于数据仓库的一个重要分析工具——联机分析处理(OLAP)的概念最早是由关系数据库之父 E.F.Codd 于 1993 年提出的。当时，Codd 认为联机事务处理(On-Line Transaction Processing，简称 OLTP)已不能满足终端

用户对数据库查询分析的需要，并且 SQL 对大数据库进行的简单查询也不能满足用户分析的需求。用户的决策分析往往需要对关系数据库进行大量的计算才能得到结果，而查询的结果并不能满足决策者提出的需求。因此，Codd 提出了多维数据库和多维分析的概念，即 OLAP。OLAP 与 OLTP 各有其特征，对二者的比较见下表 1-1。

表 1-1 OLAP 与 OLTP 特点对照表

OLTP 数据	OLAP 数据
原始数据	导出数据
细节性数据	综合性和提炼性数据
当前值数据	历史数据
可更新	不可更新，但周期性刷新
一次处理的数据量小	一次处理的数据量大
面向应用，事务驱动	面向分析，分析驱动
面向操作人员，支持日常操作	面向决策人员，支持管理需要

根据 OLAP 委员会的定义：OLAP 是使分析人员、管理人员或执行人员能够从多种角度对从原始数据中转化出来的、能够真正为用户所理解的、并真实反映企业特性的信息进行快速、一致、交互地存取，从而获得对数据的更深入了解的一类软件技术。OLAP 的目标是满足决策支持或多维环境特定的查询和报表需求，它的技术核心是“维”这个概念。多维数据模型作为数据仓库的逻辑结构被广泛的应用，OLAP 也可以说是多维数据分析工具的集合。数据仓库中的数据模型、数据组织，以及数据仓库中应该存什么样的数据，都是按照适合联机分析处理的标准来设计。

OLAP 大部分策略都是将关系型或普通数据进行多维数据存储，以便于进行分析，从而达到联机分析处理的目的。这种多维数据库，也被看作是一个超立方体，沿着各个方向存储数据，它允许用户沿着轴线方便地分析数据，分析要求统计数据中得出一个大致的范围。与主流事务型用户相关的分析形式一般有切片和切块以及上旋和下钻。在实际应用中，OLAP 常常包括对数据的相互查询，这项活动发生在经过多种途径的一系列分析，如对底层细节的进一步挖掘之后。用户对这种多维数据模型的操作比对其他数据模型的操作要容易和直观。例如，用户