学术专著

分布式数据库查询优化研究

赵宇兰／著

FENBUSHI
SHUJUKU CHAXUN
YOUHUA YANJIU

电子科技大学出版社

学术专著

分布式数据库查询优化研究

赵宇兰／著

FENBUSHI
SHUJUKU CHAXUN
YOUHUA YANJIU

电子科技大学出版社

图书在版编目（CIP）数据

分布式数据库查询优化研究 / 赵宇兰著. —成都：电子科技大学出版社，2016.7

ISBN 978-7-5647-3761-0

Ⅰ. ①分… Ⅱ. ①赵… Ⅲ. ①分布式数据库－查询优化－研究 Ⅳ. ①TP311.133.1

中国版本图书馆 CIP 数据核字（2016）第 154571 号

内 容 简 介

本书从分布式数据库查询优化的相关理论入手，首先对传统的分布式优化策略进行了分析比较，提出了采用组合寻优方法解决分布式多连接查询优化问题。在此基础上，对 MapReduce 环境下的等值连接算法效率优化、多连接效率优化、数据严重倾斜情况下的连接算法效率优化以及任意连接算法效率优化问题进行研究，并辅以实例和实验验证算法的有效性。全书集理论、技术、方法及实践于一体，具有较强的理论性和实践性，能够反映当前该领域的最新研究成果。

本书适合 IT 工程技术人员、分布式数据处理与大数据应用研究人员阅读，也可作为高等院校计算机及相关专业本科生和研究生的参考书籍。

分布式数据库查询优化研究

赵宇兰 著

出　　版：电子科技大学出版社（成都市一环路东一段 159 号电子信息产业大厦　邮编：610051）
策划编辑：杨仪玮　李燕芩
责任编辑：杨仪玮　李燕芩
主　　页：www.uestcp.com.cn
电子邮箱：uestcp@uestcp.com.cn
发　　行：新华书店经销
印　　刷：成都市新都华兴印务有限公司
成品尺寸：170 mm×240 mm　　印张 13　　字数 255 千字
版　　次：2016 年 7 月第一版
印　　次：2016 年 7 月第一次印刷
书　　号：ISBN 978-7-5647-3761-0
定　　价：42.00 元

■ 版权所有　侵权必究 ■

◆ 本社发行部电话：028-83202463；本社邮购电话：028-83201495。
◆ 本书如有缺页、破损、装订错误，请寄回印刷厂调换。

前　言

进入21世纪，随着科学技术和社会经济的发展，移动互联、社交网络、电子商务等应用极大扩展了互联网的边界和应用范围，各种数据迅速膨胀，渗透到每一个行业和业务职能领域，成为重要的生产因素。信息爆炸积累到了一个开始引发变革的程度，它不仅使全球充斥着比以往更多的数据，而且数据体量的剧增速度也在加快。根据国际数据公司（IDC）发布的2012年研究报告显示，2011年全球创造和复制的数据总量为1.8ZB，并且正在以每两年翻一番的速度快速增长；预计到2020年，全球产生的数据总量将超过40ZB，这将是地球上所有海滩上沙粒数量的57倍。海量数据的快速增长却也带来了数据存储、访问、分析等方面的巨大压力，急需以数据的海量存储与智能分析为特征的大数据处理技术来解决这些存在的问题。

分布式数据库技术是大数据技术的基础，它涵盖了从数据的海量存储、访问到分析等多方面的内容，是解决大数据问题的重要技术之一。近年来，随着分布式数据库技术的应用越来越广泛，分布式数据库数据处理的性能问题开始受到学术界和工业界的关注，其中有很多问题值得我们去探索和创新。分布式连接算法效率的优化就是研究的重点之一。

本书在深入研究和总结相关领域已有成果的基础上，围绕分布式连接查询效率优化问题开展研究工作。全书分为8章，具体编写思路如下。

第1章首先研究了分布式数据库查询处理相关理论，包括查询处理器的结构以及查询处理的基本流程。在此基础上对查询优化的概念和查询优化器理论做了相关阐述。

第2章分别从算法原理、优化过程、代价评估等方面对现有的分布式连接查询优化算法进行分析和比较。借鉴了站点依赖算法和分片复制算法各自的优势，提出了基于连接依赖信息的连接查询优化算法。

第3章对分布式多连接查询优化问题进行了深入分析，给出了多连接查询优化问题的数学模型和逻辑搜索空间，在此基础上设计了多连接代价评估模型。

第4章将一些经典的寻优算法应用于多连接查询优化过程中，设计了解决多连

接问题的动态规划算法、贪婪算法和遗传算法，通过比较分析确定使用遗传算法解决多连接问题具有相对较好的效果。

第 5 章针对多连接查询的特点，提出将侧重于全局搜索的遗传算法与侧重于局部搜索的模拟退火算法相结合的分布式多连接查询优化算法，阐述了算法的详细设计过程及实验结果分析。最后，论证算法的有效性。

第 6 章针对分布式编程架构 MapReduce 提出了基于 BloomFilter 的两表和多表等值连接算法。然后，基于磁盘 I/O 和网络 I/O 建立了等值连接算法的代价模型，用以选择基于 MapReduce 的最优等值连接效率方案。

第 7 章提出了针对数据倾斜的两表和多表等值连接算法。对于两表等值连接，优化了数据集中的一个或者几个数据严重倾斜时的连接算法效率。对于多表等值连接，采用基于 Range Partition 的方法，优化了用一轮 MapReduce 任务完成数据倾斜的多表连接算法效率。

第 8 章提出了基于两表和多表的任意连接算法。设计了如何用一轮 MapReduce 完成多表任意连接操作的算法 SEJ，进而在 SEJ 的基础上优化了多表任意连接。

另外，本书中所有实验采用的环境为 Hadoop 0.20.2，因此，附录 A 提供了 hadoop 0.20.2 的安装和配置过程。

本书凝聚了笔者长期的数据库实践经验以及研究思考的成果。在本书的编写过程中，广泛搜集了国内外相关材料，参考了一些最新的论著，并引用了部分材料，在此向其著作人表示感谢。电子科技大学出版社江明副社长为此书倾注了大量的心血，在此致以诚挚的谢意。

本书内容是作者的大胆探索和思考，仅代表个人观点。由于作者水平有限，书中错误和不妥之处在所难免，恳请广大专家、学者不吝批评指正。

作　者

2016 年 7 月

此为试读，需要完整PDF请访问：www.ertongbook.com

目 录

第 1 章　分布式数据库和查询优化理论基础

分布式数据库系统是计算机网络技术和数据库技术互相渗透和有机结合的产物，在当前信息社会中有着不可替代的作用。英国国家计算中心（United Kingdom National Computing Centre，UK-NCC）曾对分布式数据库做了分析和预测："分布式系统，特别是以分布式数据库为核心的分布式系统，将成为今后 10 年计算机科学发展的主要方向之一。"事实已经证明了这一点。目前，国际上每年都召开与分布式数据库相关、重点研究与探讨分布式数据库系统的各类问题及其解决方案的会议。

本章主要介绍分布式数据库的相关理论、分布式查询处理器的结构和处理流程，以及传统查询优化和分布式查询优化研究的内容，为后续章节的进一步阐述打下基础。

1.1　分布式数据库系统概论

1.1.1　分布式数据库概述

分布式数据库系统（Distributed Database System，DDBS）是物理上分散而逻辑上集中的数据库系统，它使用计算机网络将地理位置分散而管理和控制又需要不同程度集中的多个逻辑单元（如集中式数据库系统）连接起来，共同组成一个统一的数据库系统。由于它有着许多突出的特点，特别是其在网络中跨节点物理存储方面的优势——"既能够满足应用系统的局部控制和分散管理，又能实现整个组织的全局集中控制和高层次的系统管理；既能实现信息的灵活交流和共享访问，又便于统一管理和使用"，因此，被广泛应用在大型企业组织、公司集团、商业团体、跨地区管理机构以及军事国防等领域中。如今，分布式数据库系统已经成为当今信息技术的核心，特别是基于 C/S 计算模式的协作式分布式数据库系统，近年来已成为计算机科学领域最活跃的研究领域之一。

有关分布式数据库系统的研究始于 20 世纪 70 年代末期。美国计算机公司于 1976～1978 年设计世了界上第一个分布式数据库系统 SDD-1（System for Distributed Database）系统，并于 1979 年在 DEC-10 和 DEC-20 计算机上得以实现。随着计算机网络技术的飞速发展和广泛应用，分布式数据库系统领域的研究和开发开始活跃。

此为试读，需要完整PDF请访问：www.ertongbook.com

美国、西欧和日本等相继推出了规模宏大的 DDBS 研制计划，例如：①美国 IBM San Jose 实验室研制的 System R 系统和 R*；②德国斯图加特大学 E.J.Nuehold 教授领导研制的 POREL 系统；③美国加州大学伯克利分校研制的 INGRES 和荷兰阿姆斯特丹大学研制的扩展 INGRES；④法国 INRIA 研制的 SIRIUS-DELTA 系统和 IMAG 研究中心研制的 MICROBE 系统；⑤美国计算机公司（CCA）设计和实现的 DDM 系统。

1987 年，C.J.Date 在“Distributed Database:A Closer Look”中提出了完全的、真正的分布式数据库管理系统应遵循的 12 条规则，用于区分一个真正的、普遍意义上的分布式数据库系统与一个只能提供远程数据存取的系统。这 12 条规则已经被广泛接受：

①本地自治性（Local Autonomy）；
②不依赖于中心站点(No Reliance on Central Site)；
③可连续操作性(Continuous Operation)；
④数据位置透明性和独立性(Location Transparency and Location Independence)；
⑤数据分片独立性(Fragmentation Independence)；
⑥数据复制独立性(Replication Independence)；
⑦分布式查询处理(Distributed Query Processing)；
⑧分布式事务管理(Distributed Transaction Management)；
⑨硬件独立性(Hardware Independence)；
⑩操作系统独立性(Operating System Independence)；
⑪网络独立性(Network Independence)；
⑫数据库管理系统独立性(DBMS Independence)。

20 世纪 90 年代，分布式数据库系统进入商品化应用阶段。一些商品化的数据库系统产品，如 Oracle、IBM DB2、Sysbase、Microsoft SQL Server 等为了适应应用需要和扩大市场份额，先后提供了对分布式数据库的支持，不断推出和改进自己的分布式数据库产品。尽管分布式数据库技术发展迅速并日趋完善，但是由于它的建立环境复杂，技术实现有难度，完全遵循分布式数据库系统 12 条规则的商用系统仍难见到。

我国对分布式数据库系统的研究始于 20 世纪 80 年代初期，一些科研单位和高校先后建立和实现了几个各具特色的分布式数据库原型，其中包括中国科学院数学研究所和上海科技大学以及华东师范大学合作实现的 C-POREL、武汉大学数据库组研制的 WDDBS 和 WOODDBS、东北大学数据库组研制的 DMU/FO 系统、东南大学计算机系开发的 SUMDDB 系统、中国人民大学与知识工程研究所研制的

DOS/SELS 等，这些工作对我国分布式数据库技术的理论研究和应用开发起到了积极的推动作用。

进入 21 世纪以来，数据发生了爆炸性的增长。对于日益增长、趋于海量的数据的存储和管理，传统的分布式数据库模式已不再适用，新的分布式海量数据组织和管理方式应运而生。如 Google 设计的 BigTable 利用 Google 分布式文件系统来实现数据的分布式存储和管理，可以支持 PB 级的数据处理和上千台机器上的数据分布。此外，Apache 以 Google 的 BigTable 为原型开发了适用于非结构化存储的分布式数据库 HBase，可以应用于需要随机访问、实时读写的大数据环境中。同时，随着云计算时代的到来，分布式数据库的设计与开发也必将发生一定的变革。

1.1.2 分布式数据库系统的结构特点

与集中式数据库系统一样，分布式数据库系统是数据库系统的一种形式，它不仅仅包含分布式数据库管理系统和分布式数据库，还包含更多的实际内容。它是可运行的且按分布式数据库方式存储和维护数据，并向应用的网络环境系统提供数据和信息的分布式系统，如图 1-1 所示。

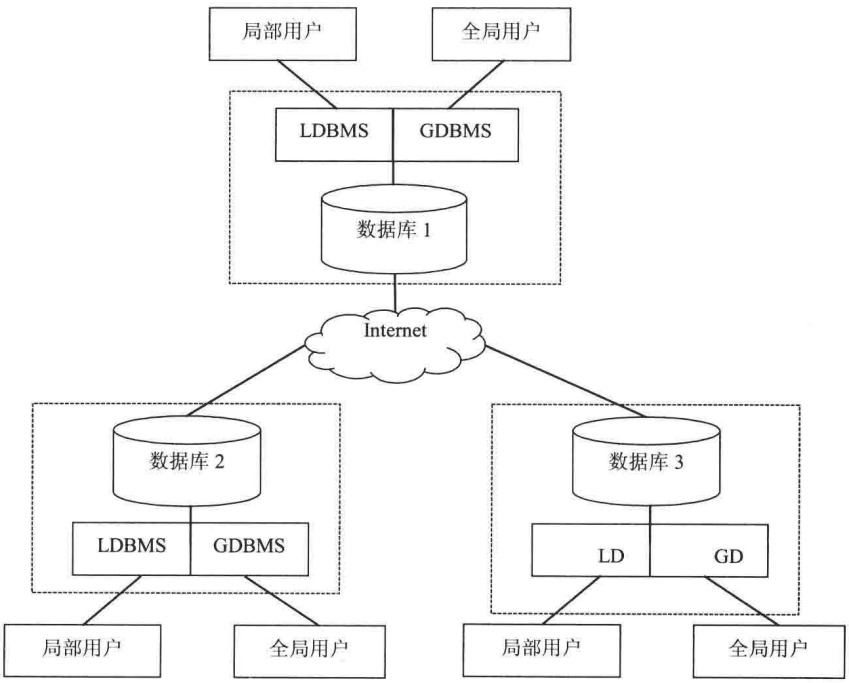

图 1-1 分布式数据库系统示意图

分布式数据库系统可以由下述部分构成。

①多台计算机设备，通过计算机网络互联。

②计算机网络设备，包括网络通信的一组软件。

③分布式数据库管理系统，包括全局数据库管理系统(GDBMS)、局部数据库管理系统（LDBMS）和通信管理程序（CM），具有全局用户接口和自治场地用户接口，并持有独立的场地目录/词典。

④分布式数据库（DDB），包括全局数据库（GDB）和局部数据库（LDB）以及自治场地数据库。

⑤分布式数据库管理者（DDBA），分为全局数据库管理者（GDBA）和局部数据库管理者（LDBA）。

⑥分布式数据库的系统软件文档，它是一组与软件相匹配的文档资料及系统的使用说明文件。

通常认为，分布式数据库系统中的数据是物理分布在用计算机网络连接起来的各个站点上；每个站点可以是一个集中式数据库系统，具有站点的局部处理能力；站点间数据相互关联，构成一个逻辑整体，共同参与并完成全局应用。由此可见，一个分布式数据库系统至少应该具备如下特点。

（1）物理分布性

物理分布性体现在数据库在网络中是跨站点物理存储的。各站点分散在不同的地方，大可为不同地区，甚至国家，小可为同一机房的不同位置。而且这种分散存储对用户透明，用户应该完全感觉不到远程与本地结合的“接缝”的存在。物理分布性是分布式数据库系统与集中式数据库系统最主要的区别之一。

（2）逻辑整体性

逻辑整体性是指分散的数据逻辑上是一个整体，它们被分布式数据库系统的所有用户共享，并由一个分布式数据库管理系统统一管理，即用户在逻辑上看到的是一个简单的、同构的、虚拟的全局数据库系统。逻辑整体性是分布式数据库系统与分散式数据库系统的最大区别。

（3）站点自治性

站点自治性是指各站点上的数据由本地的 DBMS 管理，具有自治处理能力，能够完成本地站点的局部应用，这是分布式数据库系统与多处理机系统的区别。多处理机系统虽然把数据分散存储在不同站点的数据库中，但从应用角度来看，这种数据分布与应用程序没有直接的联系，只是应用程序的执行由多个处理机并行执行，这样的系统仍然是集中式数据库系统。

（4）数据分布的透明性

在集中式数据库系统中，数据的独立性包括数据的逻辑独立性和物理独立性两个方面，用以描述应用程序与数据的逻辑结构和物理存储结构的关联性。在分布式数据库中，数据的独立性除了包括逻辑独立性与物理独立性外，还包括数据分布的独立性，即数据分布的透明性（Data Distribution Transparency）。所谓数据分布的透明性，是指用户不需要考虑数据是否分片、片段如何复制以及数据和片段如何分布等问题。数据分布的透明性是对有关数据存储方面的描述，属于数据物理独立性的范围。

（5）集中与自治相结合的控制机制

在分布式数据库中，数据的访问包含两个方面：一是局部用户共享访问本站点上的局部数据库，二是用户作为全局角色共享访问各个站点上存储的数据。因此，分布式数据库系统通常采用集中控制和自治管理相结合的机制。自治管理是指局部 DBMS 可以独立管理局部数据库，具有自治功能。同时，系统又设有集中控制机制，协调各个局部 DBMS 的工作，执行全局控制管理功能。

1.1.3 分布式数据库系统类型分析

目前对分布式数据库系统的分类还没有标准的定义，但有些提议已被认同。一种是以局部数据库的数据模型类型进行分类，另一种是按分布式数据库的全局控制系统类型进行分类。

1．按局部数据库的数据模型类型分类

（1）同构型（Homogeneous）DDBS

以构造相同的局部数据库组成的分布式数据库为同构型 DDBS。所谓构造相同，是指构成各站点局部数据库的数据模型相同。但是，具有相同数据模型的数据库若为不同公司的产品，其性质也不尽相同。因此，根据同构型 DDBS 是否采用同一厂商的 DBMS 为依据，可以将同构型 DDBS 进一步分成同构同质型和同构异质型。

早期国际上著名的同构型 DDBS 有美国 CCA 公司的 SDD-1 和 DDM、IBM 公司的 System R、德国的 POREL 和法国的 SIRIUS-DELTA 等。

（2）异构型（Heterogeneous）DDBS

如果局部站点上数据库的数据模型各不相同，则称分布式数据库为异构型 DDBS。典型的异构型 DDBS 有美国 CCA 公司的 MULTIBASE、美国佛罗里达大学的 IMDAS 和 HONEYWELL 公司的 DDTS。

除此之外，还有一些准分布式数据库系统，这些系统具备了分布式系统的部分特征，但又未能实现或未能达到分布式数据库的综合指标。典型的有 TANDEM 公

司的 ENCOMPASS 系统，IBM 公司的 CICS/ISC 系统，ORACLE 公司的 SQL*STAR、IMGRES 产品，CULLINAAE 公司的 IDMS DDS，SIEMENS AG 公司的 VDS-D，SOFTWARE AG 公司的 NET-WORK，SYBASE 公司的 REPLICATION SERVER 等。

2．按分布式数据库的全局控制系统类型分类

按照分布式数据库控制系统的全局控制类型来看，分布式数据库系统可以分类为全局控制集中型 DDBS、全局控制分散型 DDBS 和全局控制可变型 DDBS。

（1）全局控制集中型 DDBS

如果 DDBS 中的全局控制机制和全局数据字典位于一个中心站点，并由中心站点协调全局事务和控制局部数据库的转换，则称该 DDBS 为集中型 DDBS。这种方式控制机制简单，能够确保数据更新的一致性。但由于全局控制机制和全局数据字典集中存放在一个站点，该站点将成为集中失效点，一旦故障，整个系统将会崩溃。

（2）全局控制分散型 DDBS

如果 DDBS 中的全局控制机制和全局数据字典分散存储在网络的各个站点上，并且每个站点都能有效协调全局事务和控制局部数据库转换，则称该 DDBS 为分散型 DDBS。这种方式具有较好的站点独立性和高的可用性，不会因为单个站点故障而影响整个数据库运行。缺点是全局控制机制的协调和保持事务的一致性较困难，需要复杂的设施。

（3）全局控制可变型 DDBS

全局控制可变型 DDBS 介于全局控制集中型 DDBS 和全局控制分散型 DDBS 之间。这种类型的 DDBS 中，根据应用的需求，将站点分成两组：主站点组和辅站点组。主站点组中包含全局控制机制和全局数据字典（或者为一部分），辅站点组中不包含全局控制机制和全局数据字典。

1.1.4 分布式数据库系统的主要技术分析

分布式数据库系统涉及的主要技术包括分布式数据库设计、分布式查询和优化、分布式事务管理和恢复、分布式并发控制、分布式数据库的可靠性、分布式数据库的安全性等。

1．分布式数据库设计的技术和方法

由于分布式数据库存储结构的特殊性，很多集中式数据库系统的关键技术问题和组织问题在分布式数据库系统中变得更加复杂。既要考虑数据存储的本地性、并发度和可靠性，还要兼顾多副本带来同步更新的开销；既要均衡各站点的工作负荷，提高应用执行的并发度，又要考虑站点负荷分布对处理本地性的副作用。正是因为面临着种种折中考验，使分布式数据库的设计过程变得异常复杂，同时也大大增加

了优化模型的难度。

分布式数据库设计方法有两种：重构法和组合法。前者采用自顶向下的设计方法，后者采用自底向上的设计方法。重构法是指在充分理解用户应用需求的基础上，按照分布式数据库系统的设计思想和方法，一步一步构建系统的过程。这一过程包括概念设计、全局逻辑设计、分布设计、局部逻辑设计和物理设计等阶段，最后转化成与计算机系统相关的物理实现。

2．分布式查询和优化处理技术

分布式查询和集中式查询有着本质的不同。在集中式数据库中，查询优化的目的在于为每个用户查询寻求总代价最小的执行策略。通常总代价以查询处理期间的 CPU 代价和 I/O 代价之和来衡量。分布式查询除了要考虑 CPU 和 I/O 代价外，还要考虑数据在网络站点之间的传输、数据的冗余和分布对查询效率所产生的影响。分布式查询优化的两个评估标准：一种是以总代价最小为评估标准，另一种是以每个查询响应时间最短为评估标准。在实际应用中，这两种标准常常被结合使用。

为了解决分布式查询优化问题，不同的算法被提出。典型的分布式算法有基于关系代数等价变化的查询优化处理方法、基于半连接算法的查询优化处理方法和基于直接连接算法的查询优化处理方法，这些方法已被应用在实际的查询优化处理中。国内外一些学者也将现代寻优算法（如动态编程算法、贪心算法、迭代提高算法、模拟退火算法和遗传算法等）应用于分布式查询中，用来处理多连接查询优化问题，并取得了一定的效果。

3．分布式事务管理和恢复技术

集中式环境下的事务的原子性（Atomicity）、一致性（Consistency）、隔离性（Isolation）和永久性（Durability）仍然适用于分布式环境。但是与集中式相比，分布式事务管理增加了不少新的内容和复杂性。例如，多副本一致性的保证、单点故障的恢复管理问题、通信网络故障时的恢复管理问题等。为了解决这些问题，分布式事务管理程序必须同时保证本地事务的 ACID 特性和分布式事务的 ACID 特性。同时，当故障发生时，要使得分布式数据库恢复到一个正确的、一致的状态。

4．分布式并发控制技术

分布式并发控制是以集中式数据库的并发控制技术为基础，主要解决多个分布式事务对数据的并行调度问题。分布式数据库系统并发控制的主要技术包括基于分布式数据库系统并发控制的封锁技术死锁处理技术、并发控制的时标技术、并发控制的多版本一致性技术以及并发控制的乐观方法等。这些技术用于负责正确协调并发事务的执行，保证并发的存取操作不破坏数据库的完整性和一致性，确保并发执

此为试读，需要完整PDF请访问：www.ertongbook.com