MySQL Core

Broadview
www.broadview.com.cn

MySQL内核

InnoDB存储引擎

卷1

姜承尧 蒋鸿翔 饶珑辉 温正湖 著

MySQL领域Oralce ACE专家力作

深入MySQL数据库内核源码分析

InnoDB存储引擎内核开发与优化必备宝典

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

MySQL内核

InnoDB存储引擎

卷1

姜承尧 蒋鸿翔 饶珑辉 温正湖 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书由资深 MySQL 专家亲自执笔，在以往出版的两本 InnoDB 介绍性图书的基础之上，更深入地介绍 InnoDB 存储引擎的内核，例如 latch、B+树索引、事务、锁等，从源代码的角度深度解析了 InnoDB 的体系结构、实现原理、工作机制，并给出了大量最佳实践，希望通过本书帮助用户真正了解一个数据库存储引擎的开发。

本书可以成为带领读者进入数据库存储引擎内核开发的入门书籍，帮助那些从事 MySQL 数据库相关行业的从业人员。同时，本书也适合于在研究生阶段有志于进行数据库内核开发的同学阅读。

图书在版编目（CIP）数据

MySQL 内核：InnoDB 存储引擎. 第 1 卷 / 姜承尧等著. —北京：电子工业出版社，2014.5
ISBN 978-7-121-22908-4

Ⅰ. ①M… Ⅱ. ①姜… Ⅲ. ①关系数据库系统 Ⅳ.①TP311.138

中国版本图书馆 CIP 数据核字（2014）第 067306 号

策划编辑：孙学瑛
责任编辑：徐津平
特约编辑：顾慧芳
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：787×980　1/16　印张：22.5　字数：528 千字
印　　次：2014 年 5 月第 1 次印刷
定　　价：69.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

前言

为什么要写这本书

过去这些年，我一直在和各种不同的数据库打交道，见证了 MySQL 从一个小型的关系型数据库发展成为各大互联网企业的核心数据库系统的过程。期间我参与了一些大大小小的项目开发工作，成功地帮助开发人员构建了一些可靠的、健壮的应用程序。在这个过程中积累了一些经验，正是这些不断累积的经验赋予了我灵感，于是有了本书。这本书实际上反映了这些年来我做了哪些事情，汇集了很多同行每天可能都会遇到的一些问题，并给出解决方案。

本书是 MySQL 内核系列的第一本书，与之前出版的 MySQL 技术内幕不同的是，该系列的书将更靠近数据库内核层面，揭示 MySQL 数据库内核是如何运行的。MySQL 内核系列的第一本书将从 InnoDB 存储引擎的内核来展开。

毫无疑问，InnoDB 存储引擎已经成为 MySQL 数据库的“标准配置”。Facebook、Twitter、Yahoo、百度、淘宝、腾讯、网易这些互联网公司都将 InnoDB 作为后台的存储引擎。在时间的长河以及线上高并发验证下，其已经被证明是高性能、高可扩展性的引擎。

身处数据库这个圈子，可以明显地感觉到从 2010 年开始，各大互联网公司已经不再满足于仅仅使用 InnoDB 存储引擎，他们开始越来越接触到引擎的内核层面，对引擎进行内核级别的优化以及根据公司的业务需求进行二次开发。即使是 DBA 本身也开始慢慢地不满足现状开始研究起 InnoDB 存储引擎的内核，似乎一夜之间不了解点内核实现都不好意思和别人说你是搞 MySQL 数据库的。

当然，我们需要感谢 MySQL 数据库，感谢 MySQL 数据库的创始人和 InnoDB 存储引擎的创始人。正是他们开源了这些代码，使得我们这些后人可以站在巨人的肩膀上继续学习与进步。在这方面，MySQL/InnoDB 比其他数据库都要伟大，更值得我们尊敬。

不可否认的是，国内对于数据库内核的开发学习资料与课程都非常有限。本科阶段几乎没有相关课程，仅特定数据库研究方向的研究生才会去关注这些技术，并且这些人才在国内非常稀少。很多想要踏进数据库内核领域的人在最初都会感到迷茫和无助。

另外，有些人凭着自己的聪明与天赋看似掌握了内核的实现，但是从他们的博客描述来看，其离真正的理解还是有一些距离的，或者说他们仅刚入门。所以我们才会在网上看到不断有人在翻阅过代码后，或者简单设置了几个断点和调试后抱怨 InnoDB 存储引擎的设计是多么烂。

其实数据库的世界并不如他们想象的那样简单与粗糙，数据库有着自己的理论体系。虽然数据库的实现有很多种，但大多需要遵循一些理论规范，如 Fix Rules、Write-Ahead Log、Force-log-at-commit、Lock 等。

我从 2006 年就开始进行数据库的内核开发，现在想来还最多只能称为 hack。我在内核开发的路上走了很多弯路，经过高人的指点以及自己不断的学习与探索，终于有了一些经验，现通过本书来完整地展示给读者。希望通过 MySQL 内核系列，使正在通往或已经在数据库内核开发道路的人员少走弯路。

出于这个目的，我联合了网易 MySQL 技术组的各位同事，完成了 InnoDB 存储引擎卷 1 的书籍撰写工作。其中第 1、3、4、5、7、8、9、10、11 章由我个人独立完成，第 2 和第 14 章由我和温正湖共同完成、第 6 和第 12 章由我和饶珑辉共同完成、第 13 和第 15 章由我和蒋鸿翔共同完成。

在每章的最后，我还给出了思考题以及继续阅读的参考资料，通过这部分的内容，读者可以加深对于每个知识模块的理解，并继续对某一模块进行深入研究。

读者对象

本书面向的读者群：

- 数据库管理员
- 数据库架构设计师
- 数据库内核开发人员
- 其他对数据库内核感兴趣的开发人员

如何阅读本书

本书一共有 15 章，每章都像一本“迷你书”，可以单独成册。用户可以有选择地阅读，但是更推荐根据本书的组织方式进行阅读，这样会更具有条理性。

第 1 章 概览

本章首先介绍了 MySQL 数据库以及 InnoDB 存储引擎的历史，之后介绍了 InnoDB 存储引擎的源码结构与代码风格，最后推荐了阅读 InnoDB 存储引擎源码的次序。

第 2 章 基本数据结构与算法

本章对 InnoDB 中常用的数据结构和算法进行了介绍。首先是 InnoDB 的内存管理系统，从内存管理机制、内存操作基元和内存池及内存区等概念着手进行了详细讲解；之后是哈希表结构，介绍了简单哈希表和带链哈希表两种；然后介绍了双链表结构；最后还介绍了动态数组、标准排序函数。本章的内容是 InnoDB 的基础，相信读者在阅读后续章节的代码时一定会遇到本章所提的相关数据结构与算法。

第 3 章 同步机制

本章介绍了 InnoDB 存储引擎中实现的同步机制 mutex 和 rw-lock。InnoDB 存储引擎正是通过这些数据结构才能完成正确并发控制的。

第 4 章 重做日志

本章首先介绍与重做日志模块相关的概念，之后具体分析了 InnoDB 存储引擎重做日志模块的实现。InnoDB 存储引擎原先就支持组提交，因此有着相当不错的性能。最后，根据之前所介绍的内容，分析了如何通过重做日志进行有效恢复，从而实现事务系统持久性的要求。

第 5 章 mini-transaction

本章介绍了数据库中的三个协议：FIX Rules、Write-Ahead Log、Force-Log-at-commit，同时介绍了 InnoDB 存储引擎中 mini-transaction 的实现，并通过一个示例简单展示了 mini-transaction 产生的重做日志内容。

第 6 章 存储管理

本章介绍了 InnoDB 存储引擎的物理存储方式，这包括表空间的构成，段、区、页的存储管理。此外，还介绍了 InnoDB 存储引擎的文件操作方式，包括文件操作的架构设计、同步读 / 写方式和异步读 / 写方式，分别介绍了 Windows 操作系统、Posix 操作系统以及 InnoDB 模拟的三种异步 I/O 的实现方法。

第 7 章 记录

本章介绍了 InnoDB 存储引擎的记录（record），使读者了解在源码中记录可以分为物理记录与逻辑记录，以及各种记录所使用的场合。

第 8 章 索引页

本章介绍了 InnoDB 存储引擎的索引页，知道在源码中页可以分为物理页与逻辑页，并且详细分析了 page header 以及 page directory。此外，还对 InnoDB 存储引擎如何在页中进行记录的定位、插入和删除等操作进行了详细介绍。

第 9 章 锁

本章介绍了 InnoDB 存储引擎锁的实现技术。在 InnoDB 存储引擎中，其通过 next-key locking 算法在事务隔离级别 REPEATABLE READ 实现了完全的隔离性要求。此外，其对锁的设计是一种极其高效的设计方式。每个内核开发人员都应该细读 lock 模块，从而更为深入地理解锁的内部实现。

第 10 章 B+ 树索引

本章对 InnoDB 存储引擎的 B+ 树索引实现做了十分详细的介绍。该部分所需要涉及的内容非常多，与前面章节的联系也比较紧密，是一个极为重要的章节。希望读者可以反复阅读，从而更好地体会 InnoDB 存储引擎中 B+ 树索引的实现。

第 11 章 Insert Buffer

本章介绍了 InnoDB 存储引擎中 Insert Buffer 的实现，首先介绍了 Insert Buffer 的基本概念，然后介绍了 Insert Buffer 的物理与逻辑存储结构，并通过一个示例进行展示。最后，介绍了 Insert Buffer 的源码实现。我认为这个模块是难度最大的模块之一。

第 12 章 缓冲池

本章介绍了 InnoDB 存储引擎缓冲池的实现，这包括缓冲池的管理、页的读取和页的刷新。此外，还介绍了 InnoDB 存储引擎使用 midpoint insertion strategy LRU 的 LRU 管理机制。

第 13 章 事务处理

本章介绍了 InnoDB 存储引擎的事务处理模块，介绍了 InnoDB 存储引擎对于 undo 记录的存储方式，这其中涉及事务系统段、回滚段、undo 段、undo 页、undo 日志、undo 记录等多个概念，读者应该好好地理清这些概念。此外，还讲述了事务的 purge、rollback、commit 等操作的具体实现。相信通过本章的学习读者可以了解如何设计一个高效的事务系统。

第 14 章 数据字典

本章介绍了 InnoDB 存储引擎对于数据字典的具体实现，以及其与之前各章的联系。

第 15 章 服务管理

本章介绍了 InnoDB 存储引擎各服务模块的管理，并展示了这些服务模块的具体实现。

勘误和支持

由于水平有限，编写时间仓促，书中难免会出现一些错误或不准确的地方，恳请读者批评指正，

我将尽力在线上为你提供最满意的解答。如果你有更多的宝贵意见，也欢迎发送邮件至邮箱jiangchengyao@gmail.com，期待能够得到你最真挚的反馈。

致谢

感谢网易研究院的所有同事们，能与一群才华出众的人一起工作让我感到非常荣幸与自豪，同时通过不断地与他人的交流，使我在数据库方面得到了极大的提升和领悟。

感谢电子工业出版社博文视点公司的孙学瑛老师，她在这段时间内始终支持我的写作，正是她的鼓励和帮助引导我顺利完成全部书稿。

谨以此书献给我最亲爱的家人，以及众多热爱 MySQL 数据库的朋友们！

姜承尧（David Jiang）

2014 年 3 月于中国杭州

目录