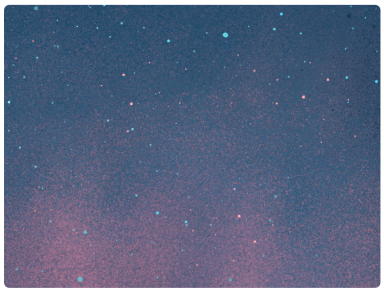


图文 33 MySQL数据库的日志顺序读写以及数据文件随机读写的原理

详情 评论

MySQL数据库的日志顺序读写以及数据文件随机读写的原理

- **如何提问**：每篇文章都有评论区，大家可以尽情留言提问，我会逐一答疑
- **如何加群**：购买狸猫技术窝专栏的小伙伴都可以加入狸猫技术交流群，一个非常纯粹的技术交流的地方

具体加群方式，请参见目录菜单下的文档：**《MySQL专栏付费用户如何加群》**（购买后可见）

之前我们花了很多篇幅去讲解MySQL的底层数据存储结构，其实那些知识是极为枯燥的，因为大部分时候，MySQL在底层如何存储数据的一些细节，比如什么数据头、附加信息之类的极为复杂，大家直接那么研究是很痛苦的。

所以我之前也就初步的给大家介绍了一下数据行、数据页、extent、extent分组、表空间、磁盘文件这些概念，主要是让大家把物理数据结构与Buffer Pool缓存的结合使用，有一个理解就行了。

掌握到之前的一些知识，基本上MySQL稍微进一步的原理，大家也就有一定的了解了。其实暂时来说这就足够了，更加细节的一些知识，比如表空间的存储结构细节，extent的存储结构细节，都要结合未来的索引优化原理、数据删除原理，结合这些东西去分析，大家从自己日常都接触的一些场景出发，去看一些技术细节，才能真正很好理解。

那么今天开始，我们将要用连续几天的时间，给大家介绍一个**真实的生产优化案例**，这个案例主要用到的知识，其实大家之前都学过了

所以这也是我一如既往的专栏风格，讲一些理论，同时插入一些我们生产环境的真实案例分析，让大家理论和实战结合起来。

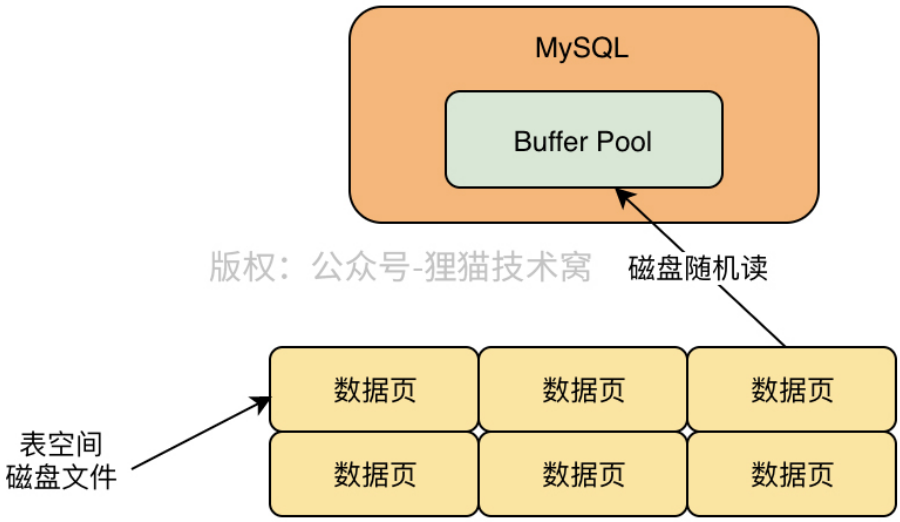
在讲解这个真实的生产案例之前，有一些前置的知识要给大家介绍一下

首先今天要讲解的就是MySQL数据库和底层的操作系统之间的交互原理，理解了这个原理后，我们再一步步剖析一个生产环境的MySQL数据库每隔一两个月性能就会出现急剧抖动的案例。

先给大家剖析一下MySQL在实际工作时候的两种数据读写机制，一种是对redo log、binlog这种日志进行的磁盘顺序读写，一种是对表空间的磁盘文件里的数据页进行的磁盘随机读写。

简单来说，MySQL在工作的时候，尤其是执行增删改操作的时候，肯定会先从表空间的磁盘文件里读取数据页出来，这个过程其实就是典型的磁盘随机读操作

我们先看下面的图，图里有一个磁盘文件的示意，里面有很多数据页，然后你可能需要在一个随机的位置读取一个数据页到缓存，这就是**磁盘随机读**



因为你要读取的这个数据页可能在磁盘的任意一个位置，所以你在读取磁盘里的数据页的时候只能是用随机读的这种方式。

磁盘随机读的性能是比较差的，所以不可能每次更新数据都进行磁盘随机读，必须是读取一个数据页之后放到Buffer Pool的缓存里去，下次要更新的时候直接更新Buffer Pool里的缓存页。

对于磁盘随机读来说，主要关注的性能指标是IOPS和响应延迟

IOPS之前给大家介绍过，就是说底层的存储系统每秒可以执行多少次磁盘读写操作，比如你底层磁盘支持每秒执行1000个磁盘随机读写操作和每秒执行200个磁盘随机读写操作，对你的数据库的性能影响其实是非常大的。

这个IOPS指标如何观察，之前也讲过了，大家在压测的时候可以观察一下。这个指标实际上对数据库的crud操作的QPS影响是非常大的，因为他在某种程度上几乎决定了你每秒能执行多少个SQL语句，底层存储的IOPS越高，你的数据库的并发能力就越高。

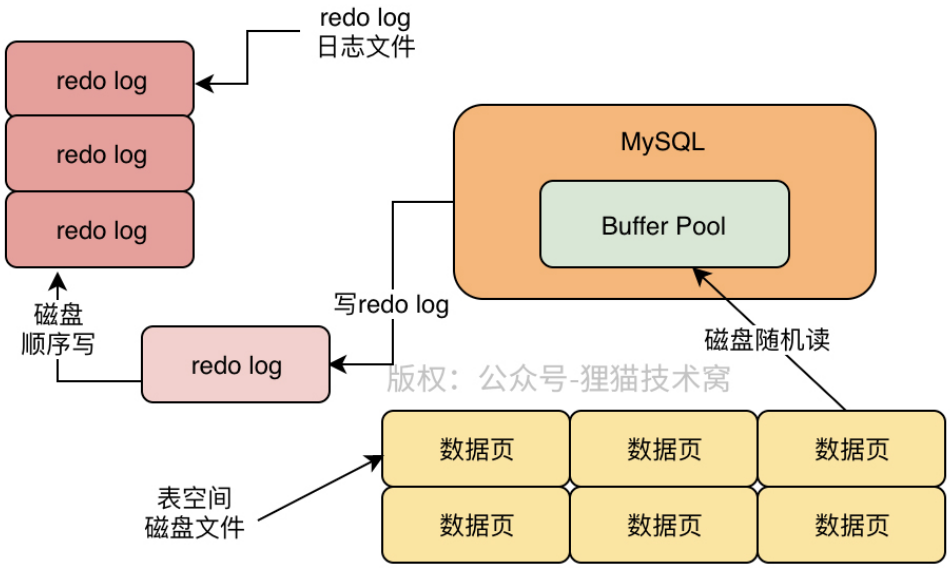
另外一个就是磁盘随机读写操作的响应延迟，也是对数据库的性能有很大的影响。因为假设你的底层磁盘支持你每秒执行200个随机读写操作，但是每个操作是耗费10ms完成呢，还是耗费1ms完成呢，这个其实也是有很大的影响的，决定了你对数据库执行的单个crud SQL语句的性能。

比如你一个SQL语句发送过去，他磁盘要执行随机读操作加载多个数据页，此时每个磁盘随机读响应时间是50ms，那么此时可能你的SQL语句要执行几百ms，但是如果每个磁盘随机读仅仅耗费10ms，可能你的SQL就执行100ms就行了。

所以其实一般对于核心业务的数据库的生产环境机器规划，我们都是推荐用SSD固态硬盘的，而不是机械硬盘，因为SSD固态硬盘的随机读写并发能力和响应延迟要比机械硬盘好的多，可以大幅度提升数据库的QPS和性能。

接着我们来看磁盘顺序读写，之前我们都知道，当你在Buffer Pool的缓存页里更新了数据之后，必须要写一条redo log日志，这个redo log日志，其实就是就是走的顺序写

所谓顺序写，就是说在一个磁盘日志文件里，一直在末尾追加日志，我们看下图。



所以上图可以清晰看到，写redo log日志的时候，其实是不停的在一个日志文件末尾追加日志的，这就是磁盘顺序写。

磁盘顺序写的性能其实是很高的，某种程度上来说，几乎可以跟内存随机读写的性能差不多，尤其是在数据库里其实也用了os cache机制，就是redo log顺序写入磁盘之前，先是进入os cache，就是操作系统管理的内存缓存里。

所以对于这个写磁盘日志文件而言，最核心关注的是磁盘每秒读写多少数据量的吞吐量指标，就是说每秒可以写入磁盘100MB数据和每秒可以写入磁盘200MB数据，对数据库的并发能力影响也是极大的。

因为数据库的每一次更新SQL语句，都必然涉及到多个磁盘随机读取数据页的操作，也会涉及到一条redo log日志文件顺序写的操作。所以磁盘读写的IOPS指标，就是每秒可以执行多少个随机读写操作，以及每秒可以读写磁盘的数据量的吞吐量指标，就是每秒可以写入多少redo log日志，整体决定了数据库的并发能力和性能。

包括你磁盘日志文件的顺序读写的响应延迟，也决定了数据库的性能，因为你写redo log日志文件越快，那么你的SQL语句性能就越高。

所以今天就先给大家在之前知识的基础之上，讲解一下数据库运行过程中，磁盘随机读写和磁盘顺序读写的两个机制