

MySql笔记-innodb无特殊说明，基于5.7版本

[零 参考资料 3](#_Toc43307726)

[一 概述 3](#_Toc43307727)

[1 3](#_Toc43307728)

[二 表的逻辑存储 3](#_Toc43307729)

[2 概述 3](#_Toc43307730)

[3 系统表空间(System tablespace) 4](#_Toc43307731)

[4 file-per-table表空间 4](#_Toc43307732)

[4.1 概述 4](#_Toc43307733)

[4.2 优势 4](#_Toc43307734)

[4.3 劣势 5](#_Toc43307735)

[5 一般表空间(General Tablespaces) 5](#_Toc43307736)

[5.1 概述 5](#_Toc43307737)

[5.2 限制 5](#_Toc43307738)

[6 Undo Tablespaces 6](#_Toc43307739)

[7 临时表空间(The Temporary Tablespace) 6](#_Toc43307740)

[8 Segment 6](#_Toc43307741)

[9 Extent 6](#_Toc43307742)

[10 Page 6](#_Toc43307743)

[10.1 概述 6](#_Toc43307744)

[10.2 File Header 7](#_Toc43307745)

[10.3 Page Header 7](#_Toc43307746)

[10.4 Infimun+Supremum Records 7](#_Toc43307747)

[10.5 User Records 7](#_Toc43307748)

[10.6 Free Space 7](#_Toc43307749)

[10.7 Page Directory 7](#_Toc43307750)

[10.8 File Trailer 7](#_Toc43307751)

[10.9 页文件分析 7](#_Toc43307752)

[11 Row 8](#_Toc43307753)

[11.1 概述 8](#_Toc43307754)

[11.2 REDUNDANT 8](#_Toc43307755)

[11.3 COMPACT 9](#_Toc43307756)

[11.4 DYNAMIC 10](#_Toc43307757)

[11.5 COMPRESSED 10](#_Toc43307758)

[11.6 COMPACT 行格式文件解析 10](#_Toc43307759)

[12 文件格式 10](#_Toc43307760)

[二 索引 11](#_Toc43307761)

[1 概述 11](#_Toc43307762)

[2 B树 11](#_Toc43307763)

[3 B+树 11](#_Toc43307764)

[4 Clustered index 11](#_Toc43307765)

[5 Secondary index 11](#_Toc43307766)

[6 哈希索引(Hash index) 11](#_Toc43307767)

[7 覆盖索引(Covering Index) 11](#_Toc43307768)

[8 全文检索(Full-Text Search)和倒排索引(Inverted Index) 11](#_Toc43307769)

[9 空间索引(Spatial Index)和R树(R-tree) 11](#_Toc43307770)

[三 InnoDB In-Memory Structures 11](#_Toc43307771)

[1 LUR(the least recently used) 算法 11](#_Toc43307772)

[2 Buffer Pool 11](#_Toc43307773)

[2.1 概述 11](#_Toc43307774)

[2.2 缓冲池配置 12](#_Toc43307775)

[2.3 控缓冲池监控参数 12](#_Toc43307776)

[3 Change Buffer 12](#_Toc43307777)

[3.1 概述 12](#_Toc43307778)

[3.2 相关配置 13](#_Toc43307779)

[3.3 相关监控 13](#_Toc43307780)

[4 Log Buffer 13](#_Toc43307781)

[5 自适应哈希索引(Adaptive Hash Index) 13](#_Toc43307782)

[四 锁(Lock) 14](#_Toc43307783)

[五 日志(Log) 14](#_Toc43307784)

[六 事务(Transaction) 14](#_Toc43307785)

[七 其他 14](#_Toc43307786)

[1 、配置 14](#_Toc43307787)

[2 、优化 14](#_Toc43307788)

[3 、限制 14](#_Toc43307789)

# 零 参考资料

《Mysql技术内幕 InnoDB存储引擎》

《高性能Mysql 第三版》

《高可用Mysql》

《算法导论 第三版》

《漫画算法：小灰的算法之旅》

《算法 第四版》

<https://dev.mysql.com/doc/refman/5.7/en/>

<https://dev.mysql.com/doc/refman/5.7/en/innodb-on-disk-structures.html>

# 概述

* + - 2. 整体架构图



# 表的逻辑存储

## 概述

* + - 1. 索引组织表(index organized table)：表都是根据主键顺序存放
      2. 如果没有显式指定主键，则：
         1. 判断表中是否由唯一非空索引，有则指定其为主键，如果有多个，则根据定义索引的顺序，指定第一个为主键
         2. 如果b.1不符合，则自建一个6字节大小的指针键
      3. 单列主键情况下，可以用\_rowid查看主键情况。
      4. 所有数据都被逻辑地放在表空间中(tablespace),表空间又由段(segment),区(Extent),页(page,或者称为块(block))
      5. 逻辑存储结构



## IO和文件管理

## 文件格式

### 概述

* + - 1. Innodb目前支持两种文件格式：Antelope、Barracuda
         1. Antelope

是最初的文件格式，之前甚至没有名字。。。

仅支持COMPACT和REDUNDANT行格式

## 系统表空间(System tablespace)

<https://dev.mysql.com/doc/refman/5.7/en/innodb-system-tablespace.html>

* + - 1. 可以有多个表空间数据文件，默认情况下，为一个名为ibdata1的在数据目录下的初始大小为10M的单文件，可以通过innodb\_data\_file\_path参数配置存储了Innodb的数据字典(the InnoDB data dictionary)，双写缓冲(the doublewrite buffer)，改变缓冲(the change buffer)，undo日志(undo logs)以及一些创建在此空间的表的
      2. 当innodb\_file\_per\_table参数禁用且未指定表空间时，默认在此空间创建表
      3. 当操作引起表空间数据减少时，不会收缩表空间大小，只是将释放出来的空间标记为可用

## file-per-table表空间

<https://dev.mysql.com/doc/refman/5.7/en/innodb-file-per-table-tablespaces.html>

### 概述

* + - 1. 当innodb\_file\_per\_table启用且未指定表空间时，会为每个创建表动作在数据目录下独立创建一个对应的文件名为${表名}.idb的表空间文件

### 

### 优势

* + - 1. 在truncating 或 dropping操作后，可以将存储空间还给系统
      2. 表复制ALTER TABLE操作，不会增加存储空间占用
      3. TRUNCATE TABLE操作性能相对更好
      4. 此表空间文件可以创建在不同的存储空间中
      5. 可以从其他Mysql直接迁移指定的表
      6. 可以使用Barracuda文件格式
      7. 当数据异常时可以更快速修复以及提高恢复概率
      8. 在不打断其他表使用的情况下，可以迅速恢复/备份表
      9. 可以通过监视表空间文件大小，监视文件系统上的表大小
      10. 当设置innodb\_flush\_method值为O\_DIRECT时，可以提高性能
      11. 每个表都是64TB大小极限(Tables in a shared tablespace are limited in size by the 64TB tablespace size limit. By comparison, each file-per-table tablespace has a 64TB size limit, which provides plenty of room for individual tables to grow in size.)

### 劣势

* + - 1. 此表空间只能被指定表使用，导致资源浪费
      2. 会导致有很多fsync操作
      3. mysqld必须保存很多文件句柄，导致影响性能
      4. 每个表需要更多文件描述符
      5. 如果没管理好，会产生很多碎片，影响dropTable和表扫描操作。反之会提高性能
         1. 因为缓冲池扫描会用到一个内部锁，所以在此模式下，会延迟其他操作执行
         2. 配置innodb\_autoextend\_increment只应用于共享表空间文件，这个空间固定每次扩展4MB

## 一般表空间(General Tablespaces)

https://dev.mysql.com/doc/refman/5.7/en/general-tablespaces.html

### 概述

* + - 1. 此空间是个通过CREATE TABLESPACE创建，可共享的表空间，文件后缀为.isl
      2. 存储表空间元数据消耗内存比file-per-table tablespaces少
      3. 表空间文件可以存在相对目录或独立的数据目录
      4. 支持Antelop和Barracuda文件格式，因此支持所有行格式和相关功能
      5. 参数innodb\_file\_format 和 innodb\_file\_per\_table不影响此表空间
      6. 在执行Create table时，可以指定tablespace
      7. 在执行Alter table时，可以通过Tablespace，对表做表空间迁移
      8. 由于物理页不同，压缩和非压缩的表不能在同一个表空间共存
      9. 表空间如果包含压缩表，FILE\_BLOCK\_SIZE必须指定具体值，同时innodb\_page\_size不能为32k或64k,而且KEY\_BLOCK\_SIZE必须等于FILE\_BLOCK\_SIZE/1024
      10. 当表空间中最后一张表被删除时，不会自动删除表空间文件，必须执行drop tablespace
      11. 此表空间不属于特定database，drop database操作能删掉此空间的表，但删不掉此空间
      12. 和系统空间一样，修改表数据不会改变空间大小，但会在表空间内部释放资源
      13. 表空间创建语句

1. **CREATE** TABLESPACE tablespace\_name
2. **ADD** DATAFILE 'file\_name'
3. [FILE\_BLOCK\_SIZE = value]
4. [ENGINE [=] engine\_name]

### 限制

* + - 1. 不能将其他已存在的表空间改成General tablespace
      2. 不支持创建临时General tablespace
      3. Generaltable不支持临时表
      4. 此空间中的表不支持DISCARD TABLESPACE 和IMPORT TABLESPACE操作
      5. MySQL 5.7.24以后会不支持此表空间中的表分区
      6. 不支持同主机主从复制

## Undo Tablespaces

https://dev.mysql.com/doc/refman/5.7/en/innodb-undo-tablespaces.html

## 临时表空间(The Temporary Tablespace)

https://dev.mysql.com/doc/refman/5.7/en/innodb-temporary-tablespace.html

## Segment

* + - 1. 在Innodb中，段的管理由引擎自动完成
      2. 这里的段跟回滚段不同
      3. 如果表空间是目录，段就是目录中文件
      4. 段每次扩容都一次1个区，每次释放也是一次一个区
      5. file-per-table 表空间，数据都在一个段中，对应的索引在各自的段中
      6. system tablespace表空间则包含各种不同的段
      7. 在Innodb中每个索引分配两个段，一个用于Btree的非叶子节点，一个用于叶子节点
      8. 个别Extent会分配在多个Segment中

## Extent

* + - 1. 由连续页组成的空间，
         1. 当页大小为4KB、8KB、16KB时，区大小固定为1MB；
         2. 当页大小为32KB，对应的区大小为2MB，
         3. 页大小为64KB,对应的区大小为4MB
      2. 默认情况下，一个区包含64个大小为16KB的page

## Page

### 概述

* + - 1. 结构：File Header，Page Header，Infimun和Supremum Records，User Records，Free Space，Page Directory，File Trailer



### File Header

### Page Header

### Infimun+Supremum Records



### User Records

### Free Space

### Page Directory

### File Trailer

### 页文件分析

## Row

### 概述

* + - 1. 行格式(row format): 决定了表的每行物理存储格式，直接影响到查询和DML操作的性能
      2. 硬盘上单页中存储的行更多，性能越好
      3. 行记录数据格式有四种

| **格式** | **紧凑存储特征** | **Enhanced Variable-Length Column Storage** | **Large Index Key Prefix Support** | **支持压缩** | **表空间类型** | **文件格式** |
| --- | --- | --- | --- | --- | --- | --- |
| REDUNDANT | No | No | No | No | system  file-per-table, general | Antelope  Barracuda |
| COMPACT | Yes | No | No | No | system  file-per-table  general | Antelope  Barracuda |
| DYNAMIC | Yes | Yes | Yes | No | system  file-per-table general | Barracuda |
| COMPRESSED | Yes | Yes | Yes | Yes | file-per-table general | Barracuda |

* + - 1. 可变长列(Variable-length columns):
         1. 一种存储在B树索引节点的特例
         2. 被存储在overflow pages上(Variable-length columns that are too long to fit on a B-tree page are stored on separately allocated disk pages called overflow pages.)
         3. 这些列也称为off-page
         4. 列上的数据存储在overflow page的单链表上
         5. 根据列长度不同，为避免浪费存储，列上的值的全部或部分前缀存储在B树上(Depending on column length, all or a prefix of variable-length column values are stored in the B-tree to avoid wasting storage and having to read a separate page)

### REDUNDANT

* + - 1. 老版本Mysql采用的数据结构，5.7版本为兼容老版本而依旧存在
      2. 支持Antelope和Barracuda文件
      3. 对于可变长列，值的前768 bytes的内容存储在B树节点的索引中，其余的存储在overflow page中
      4. 对于固定长度的列，如果值实际长度大于等于768 bytes ，会被转为可变长列存储
      5. 当列的值实际长度小于768 bytes，则为节省I/O开支，不会存储在overflow page，都存储在B树节点上
      6. 特征
         1. 每个索引记录包含6 byte的头，用于将连续的记录连在一起，以及行级锁
         2. 聚集索引的记录包含所有用户定义的列，一个6 byte的事务ID字段，一个7 byte的滚动指针(roll pointer)字段
         3. 如果用户没有定义表的主键，聚集索引还包括一个6 byte的row id字段
         4. 每个辅助索引(Secondary Index)包含聚集索引中的所有不在辅助索引中的主键字段
         5. 每个记录都包含一个指向所有字段的指针。

如果所有字段总长度小于128byte，则指针大小为1byte

如果所有字段总长度大于等于128byte，则指针大小为2byte

这类指针组被称为记录目录(The array of pointers is called the record directory)

指针指向的空间，是记录的数据部分

* + - * 1. 对于数据NULL值

无论可变长度列或固定长度列，都会在记录目录(record directory)预留1byte或2byte

对于可变长度列，NULL值不会在数据区预留空间

对于固定长度列，NULL值也会预留固定长度的空间，目的是为了字段改为非NULL时不产生碎片

* + - * 1. 对于采用可变长度字符集的char(N)列，占用空间=最大字符长度\*N

### COMPACT

* + - 1. 相对于REDUNDANT格式，以增加CPU消耗为代价，减少了20%存储空间
      2. 支持Antelope和Barracuda文件
      3. 在I/O密集型环境中，性能会有一定提升；在计算密集型，性能反而下降
      4. 对于可变长列，值的前768 bytes的内容存储在B树节点的索引中，其余的存储在overflow page中
      5. 对于固定长度的列，如果值实际长度大于等于768 bytes ，会被转为可变长列存储
      6. 当列的值实际长度小于768 bytes，则为节省I/O开支，不会存储在overflow page，都存储在B树节点上
      7. 特征
         1. 每个索引记录包含5 byte头，用于将连续的记录连在一起，以及行级锁，在其前面可能还有个可变长度的头部
         2. 记录头的可变长部分(The variable-length part of the record header)

包含 1个 位矢量(bit vector)用于表明有NULL列，

如果NULL列数量为N，则这个位矢量占CEILING(N/8) bytes

除了这个位矢量, NULL值不占用其他空间

也包含可变列的长度，根据可变长列的最大长度，占1byte或2byte

如果所有列都是固定长度且NOT NULL，记录头就没有可变长部分

* + - * 1. 对于每个 非NULL的可变长度字段

记录头用1byte或2byte表示列长度

仅当当前记录需要存储在overflow page或字段最大长度超过255byte且实际长度超过127byte时，占用2byte（Two bytes are only needed if part of the column is stored externally in overflow pages or the maximum length exceeds 255 bytes and the actual length exceeds 127 bytes）

对于大字段，2 byte 表示了内部存储部分长度加上20 byte 指向外部存储的部分的指针 的长度；例如：内部数据长度时768 bytes，所以 2byte表示 768 + 20

* + - 1. 记录头后面是非NULL列的数据
      2. 聚集索引的记录包含所有用户定义的列，一个6 byte的事务ID字段，一个7 byte的滚动指针(roll pointer)字段
      3. 如果用户没有定义表的主键，聚集索引还包括一个6 byte的row id字段
      4. 每个辅助索引(Secondary Index)包含聚集索引中的所有不在辅助索引中的主键字段。如果部分主键列是可变长度的，每个辅助索引的头部都会有一个记录主键列长度的可变长度部分，即使辅助索引字列是固定长度
      5. 在内部，对于非可变长度字符集(nonvariable-length character sets)
         1. 固定长度字符列采用固定长度格式存储
         2. 尾部的空格不会从varchar列截掉
      6. 在内部，对于可变长度字符集(utf8mb3、utf8mb4)
         1. 针对char(N)

先尝试通过截断尾部空格，达到用N bytes空间存储的目的

如果实际值超过了N bytes，尾部空格将被截取到列值的最小字节长度

char(N)最大长度 = 单字符最大长度 \* N

至少要预留Nbytes空间，用于保证列完成更新并不产生碎片

### DYNAMIC

* + - 1. 整体特征跟COMPACT一样
      2. 增加了对超长可变长度列(long variable-length columns)的支持
      3. 增加了对大索引key前缀(large index key prefixes)的支持
      4. 支持Barracuda文件
      5. 当建表时指定ROW\_FORMAT=DYNAMIC，超长可变长度列的数据将完全存储在overflow page，而聚簇索引记录仅包含20byte的指向overflow page的指针。
      6. 列是否存储在off-page 依赖于页大小和行的总大小，当行太长时，最大的列将存在off-page，直到聚簇索引记录适合B树页
      7. 当TEXT 和 BLOB 列长度小于等于40byte时，存储在行中
      8. DYNAMIC格式规避了大数据量的列填满B树节点的问题
      9. 短的列将被保留在B树节点，从而最小化每行的 overflow page数量
      10. 支持索引key前缀3072 byte，需要通过innodb\_large\_prefix配置

### COMPRESSED

* + - 1. 整体特征跟COMPACT和 DYNAMIC一样
      2. 增加了支持对表和索引数据的压缩
      3. 支持Barracuda文件
      4. 采用了于DYNAMIC类似的内部细节存储off-page，同时也考虑到了因为表和索引数据被压缩以及用更小的页所带来的额外存储和性能消耗
      5. 不支持系统表空间
      6. 不支持同个表空间中同时存在压缩和不压缩的表

### COMPACT 行格式文件解析

# 索引

## 概述

## B树

## B+树

## Clustered index

## Secondary index

## 哈希索引(Hash index)

## 覆盖索引(Covering Index)

## 全文检索(Full-Text Search)和倒排索引(Inverted Index)

## 空间索引(Spatial Index)和R树(R-tree)

# InnoDB In-Memory Structures

## LUR(the least recently used) 算法

## Buffer Pool

<https://dev.mysql.com/doc/refman/5.7/en/innodb-buffer-pool.html>

### 概述

* + - 1. 实现算法是LUR算法的一个变种，默认操作流程：
         1. 将缓冲池分为两部分：
         2. 年轻子链表(new/young sublist)：占5/8缓存池容量，最新访问的页
         3. 年老子链表(old sublist)：占3/8缓存池容量，非最近访问的页，淘汰页时候选页
         4. 设置一个边界midpoint：在年轻子链表的表尾链接年老子链的表头位置
         5. 当因为用户查询操作或作为innodb自动执行的预读操作的一部分，需要读取一个新页进入缓冲池时，将新页插入到midpoint位置
         6. 当缓冲池中的一个页被存取
         7. 如果是用户操作引起，在第一次存取时，立即将此页放入到年轻子链表的头部
         8. 如果是因为innodb自动执行的预读操作引起，在第一次存取时，不会将此页放入到年轻子链表的头部，可能直到此页被淘汰，都不会被放入到年轻子链表的头部
         9. 随着数据库操作，缓冲池中的页会被逐步移向缓冲池尾部(年老子链表尾部)
         10. 每个页随着其他页被移到年轻子链表的头部而逐渐老化
         11. 年老子链中的页，同时也因为新页插入到midpoint而加速老化
         12. 最后，缓冲池中始终未被使用的页因为到达了年老子链表的尾部而被淘汰



### 缓冲池配置

### 控缓冲池监控参数

* + - 1. 执行SHOW ENGINE INNODB STATUS 命令查看

## Change Buffer

<https://dev.mysql.com/doc/refman/5.7/en/innodb-change-buffer.html>

<https://dev.mysql.com/doc/refman/5.7/en/faqs-innodb-change-buffer.html>

### 概述

* + - 1. 这部分用于缓存不在缓冲池中的secondary index 页的变更
      2. 这部分缓存会根据相关的读操作，定期合并到Buffer Pool中去，避免频繁的磁盘随机I/O
      3. 当系统有大量空闲时间或处于slow shutdown时，会对这部分缓冲执行清理操作，将数据索引页更新序列化地写入到磁盘
      4. 当涉及很多行数据和secondary index更新，合并操作可能持续几小时。
      5. 当事务提交后也可能会触发合并操作
      6. 在memory中，Change Buffer占用buffer pool一部分空间；在disk中，Change Buffer是系统表空间的一部分
      7. Change Buffer不支持包含降序索引列的索引



### 相关配置



### 相关监控

## Log Buffer

<https://dev.mysql.com/doc/refman/5.7/en/innodb-redo-log-buffer.html>

* + - 1. 用于缓冲日志数据，并定期写入到disk
      2. 通过innodb\_log\_buffer\_size 配置，默认16MB
      3. 大的Log Buffer可以使执行大事务而不需要在事务提交前将redo日志数据写入到disk中，减少disk I/O
      4. 参数innodb\_flush\_log\_at\_trx\_commit 控制log buffer 写disk方式，参数innodb\_flush\_log\_at\_timeout控制写频率

## 自适应哈希索引(Adaptive Hash Index)

<https://dev.mysql.com/doc/refman/5.7/en/innodb-adaptive-hash.html>



# Lock与latch

## 概述

## Lock类型

### Shared and Exclusive Locks

* + - 1. shared lock：允许事务读取一行数据
      2. exclusive lock：允许事务更新/删除一行

### Intention Locks

### Record Locks

### Gap Locks

### Next-Key Locks

### Insert Intention Locks

### AUTO-INC Locks

### Predicate Locks for Spatial Index

## Phantom Row

## 死锁

# 日志(Log)

# 事务(Transaction)

# 其他

## 、配置

## 、优化

## 、限制