Part 7 InnoDB数据存储结构

1. 数据库的存储结构:页

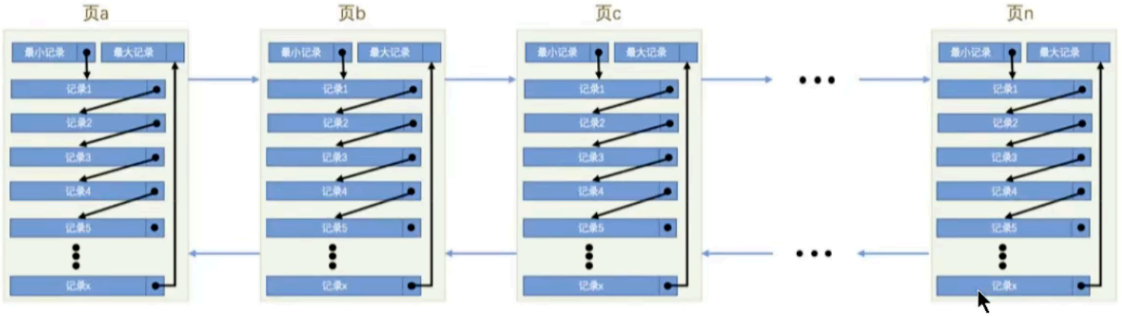
索引信息以及数据记录都是保存在文件上的，确切说是存储在页结构中。另一方面，索引是在存储引擎中实现的，MySQL服务器上的**存储引擎**负责对表中数据的读取和写入工作。不同存储引擎中**存放的格式**一般是不同的，甚至有的存储引擎比如Memory都不用磁盘来存储数据。

1.1 磁盘与内存交互基本单位:页

InnoDB将数据划分为若干个页，InnoDB中页的大小默认为 **16KB**。

数据库I/O操作的最小单位是页。一个页中可以存储多个行记录

> 记录是按照行来存储的，但是数据库的读取并不以行为单位，否则一次读取（也就是一次I/O操作）只能处理一行数据，效率会非常低。



1.2 页结构概述

页之间`双向链表`相关联。每个数据页中的记录会按照主键值从小到大的顺序组成一个`单向链表`，每个数据页都会为存储在它里边的记录生成一个`页目录`，在通过主键查找某条记录的时候可以在页目录`中使用二分法`快速定位到对应的槽，然后再遍历该槽对应分组中的记录即可快速找到指定的记录。

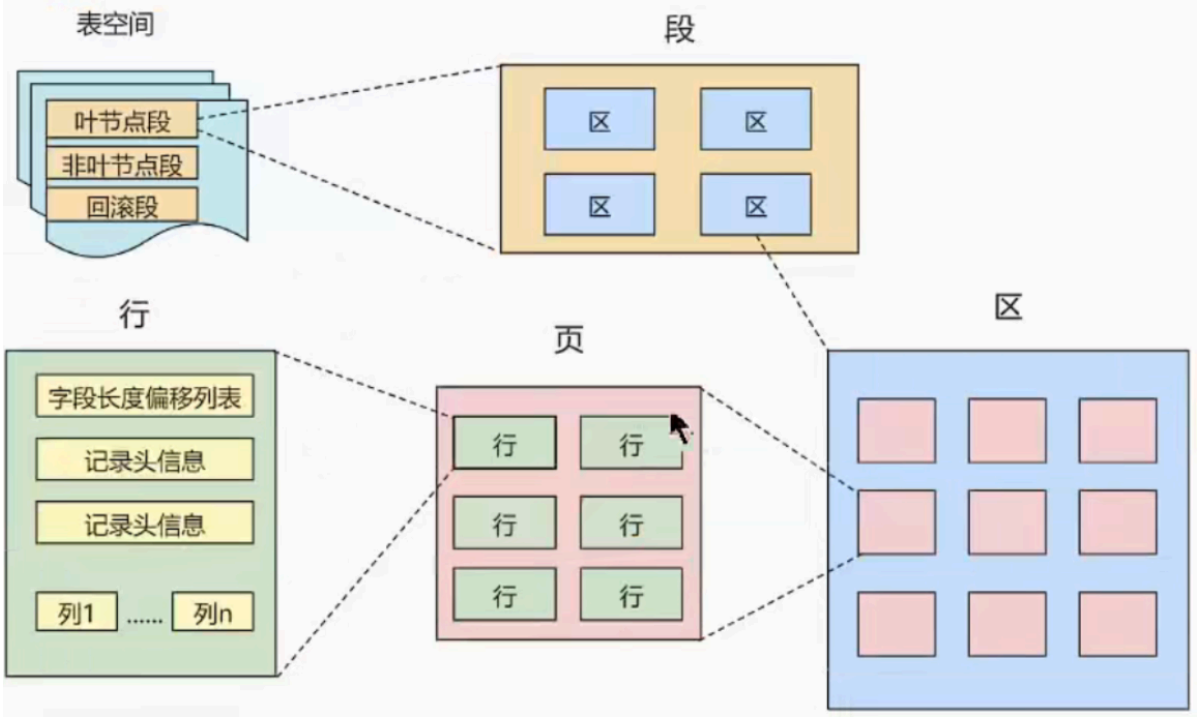
1.3 页的大小

不同的数据库管理系统（简称DBMS）的页大小不同。比如在MySQL的InnoDB存储引擎中，默认页的大小是**16KB**。可以通过`show variables like '%innodb_page_size%';`查看。

SQL Server中页的大小为`8KB`，而在oracle中用术语"`块`"(Block)来代表"页”，Oralce支持的块大小为2KB，4KB，8KB，16K8，32KB和64KB。

1.4 页的上层结构

另外在数据库中，还存在区（Extent)、段(Segment)和表空间（Tablespace)的概念。行、页、区、段、表空间的关系如下图所示：



- 区（Extent）
  - 在InnoDB存储引擎中，一个区会分配`64个连续的页`，大小为1MB。
  - 区在文件系统是一个连续分配的空间
- 段（Segment）
  - 段是数据库中的分配单位，不同类型的数据库对象以不同的段形式存在。
  - 由一个区或多个区组成
  - 当创建数据表、索引的时候，就会相应创建对应的段，比如创建一张表时会创建一个表段，创建一个索引时会创建一个索引段。
  - 在段中不要求区与区之间是相邻的
- 表空间（Tablespace）
  - 逻辑容器
  - 由一个段或多个段组成
  - 数据库由一个或多个表空间组成，表空间从管理上可以划分为系统表空间，`用户表空间`、`撤销表空间`、`临时表空间`等。

2. 页的内部结构

页如果按类型划分的话，常见的有`数据页（保存B+树节点）`、`系统页`、`Undo页`和`事务数据页`等。数据页是我们最常使用的页。

数据页的`16KB`大小的存储空间被划分为七个部分

- 文件头(File Header)：描述页的信息
- 页头(Page Header)：页的状态信息
- 最大最小记录(Infimum+supremum)：两个虚拟的行记录
- 用户记录(User Records)：存储行记录内容
- 空闲空间(Free Space)：页中还没有被使用的空间

- 页目录(Page Directory)：存储用户记录的相对位置
- 文件尾(File Tailer)： 校验页是否完整

2.1 File Header(文件头部）和File Trailer (文件尾部)

2.1.1 文件头部

描述各种页的通用信息。（比如页的编号、其上一页、下一页是谁等）

名称	占用空间大小	描述
FIL_PAGE_SPACE_OR_CHKSUM	4 字节	页的校验和（checksum值）
FIL_PAGE_OFFSET	4 字节	页号
FIL_PAGE_PREV	4 字节	上一个页的页号
FIL_PAGE_NEXT	4 字节	下一个页的页号
FIL_PAGE_LSN	8 字节	页面被最后修改时对应的日志序列位置
FIL_PAGE_TYPE	2 字节	该页的类型
FIL_PAGE_FILE_FLUSH_LSN	8 字节	仅在系统表空间的一个页中定义，代表文件至少被刷新到了对应的LSN值
FIL_PAGE_ARCH_LOG_NO_OR_SPACE_ID	4 字节	页属于哪个表空间

- `FIL_PAGE_OFFSET（4字节）`：每一个页都有一个单独的页号，InnoDB通过页号可以唯一定位一个页。
- `FIL_PAGE_TYPE（2字节）`：这个代表当前页的类型。

类型名称	十六进制	描述
FIL_PAGE_TYPE_ALLOCATED	0x0000	最新分配，还没有使用
FIL_PAGE_UNDO_LOG	0x0002	Undo日志页

类型名称	十六进制	描述
FIL_PAGE_INODE	0x0003	段信息节点
FIL_PAGE_IBUF_FREE_LIST	0x0004	Insert Buffer空闲列表
FIL_PAGE_IBUF_BITMAP	0x0005	Insert Buffer位图
FIL_PAGE_TYPE_SYS	0x0006	系统页
FIL_PAGE_TYPE_TRX_SYS	0x0007	事务系统数据
FIL_PAGE_TYPE_FSP_HDR	0x0008	表空间头部信息
FIL_PAGE_TYPE_XDES	0x0009	扩展描述页
FIL_PAGE_TYPE_BLOB	0x000A	溢出页
FIL_PAGE_INDEX	0x45BF	索引页，也就是我们所说的数据页

- `FIL_PAGE_PREV`（4字节）和`FIL_PAGE_NEXT`（4字节）：通过建立一个双向链表把许许多多的页就都串联起来了，保证这些页之间**不需要是物理上的连续，而是逻辑上的连续。**
- `FIL_PAGE_SPACE_OR_CHKSUM`（4字节）：代表当前页面的校验和（checksum）。文件头部和文件尾部都有该属性。
    - 通过对比文件头和文件尾的校验和判断该页是否完整。
- `FIL_PAGE_LSN`（8字节）：页面被最后修改时对应的日志序列位置（英文名是：Log Sequence Number）

2.1.2 File Trailer（文件尾部）（8字节）

- 前4个字节：页的校验和
- 后4个字节：页面被最后修改时对应的日志序列位置（LSN）（这个部分也是为了校验页的完整性的，如果首部和尾部的LSN值校验不成功的话，就说明同步过程出现了问题）

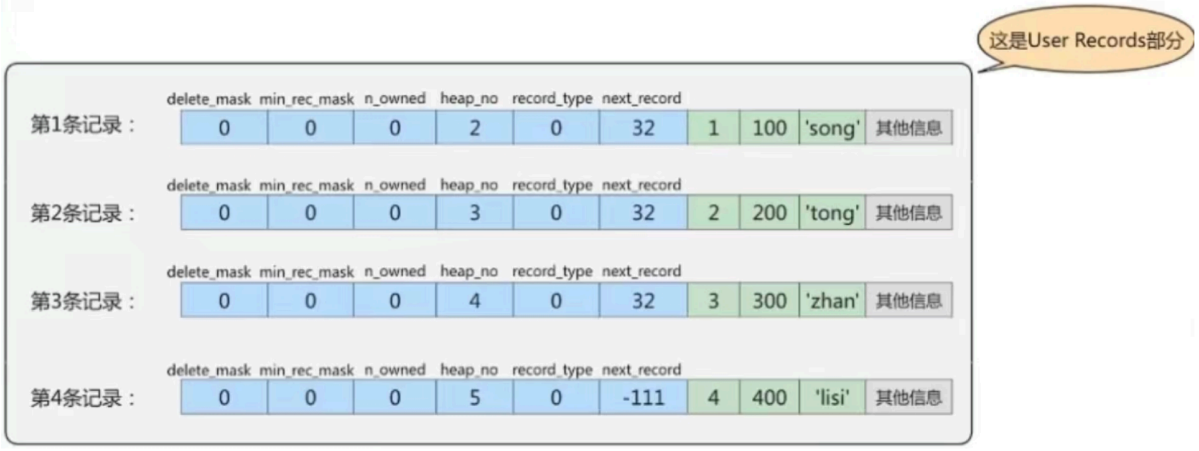
2.2 空闲空间、用户记录和最小最大记录

2.2.1 Free Space (空闲空间)

我们自己存储的记录会按照指定的`行格式`存储到`User Records`部分。每当我们插入一条记录，都会从Free Space部分，也就是尚未使用的存储空间中申请一个记录大小的空间划分到User Records部分。用完了需要去申请新页。

2.2.2 User Records (用户记录)

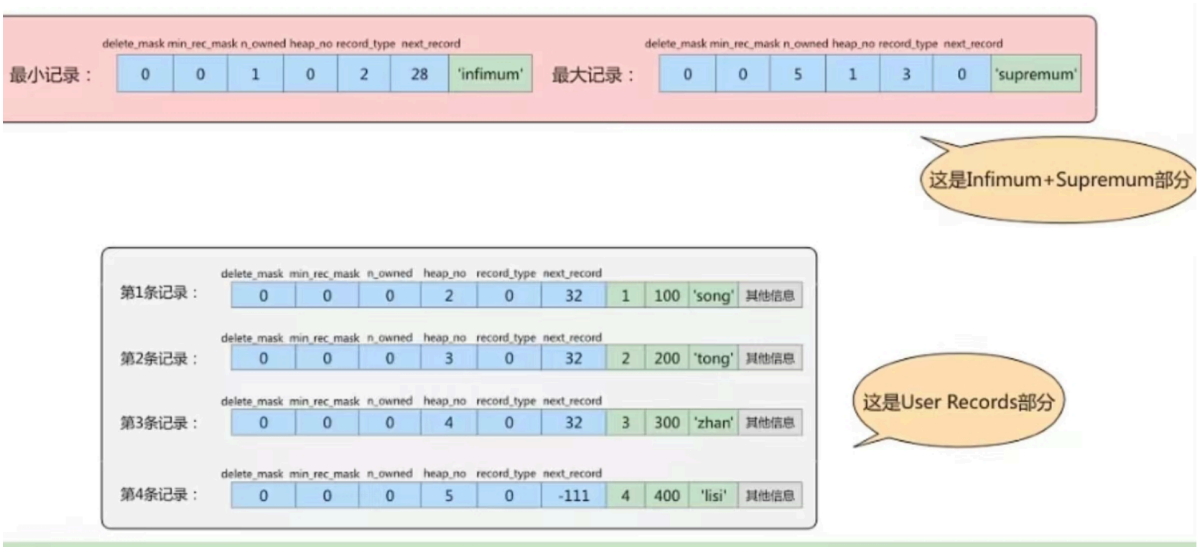
User Records中的这些记录按照`指定的行格式`一条一条摆在User Records部分，相互之间形成`单链表`。



- delete_mask：记录是否被删除。不直接删除的原因是直接删除记录效率不如修改标记的效率。
- min_rec_mask：B+树的每层非叶子节点中的最小记录都会添加该标记，值为1。
- record_type：表示当前记录的类型
  - 0为普通记录
  - 1为B+树非叶节点记录
  - 2为最小记录
  - 3为最大记录
- heap_no：当前记录在本页的位置。
  - 由于Mysql会自动给每个页加两条记录，称为伪记录或虚拟记录，一个代表最大记录，一个代表最小记录，最大记录和最小记录的heap_no值分别为0和1，所以他们的位置最靠前。
- n_owned：页目录中每个组中最后一条记录的头信息中会存储该组一共有多少条记录，作为n_owned字段。
- next_record：表示从当前记录真实数据到下一记录的真实数据的地址偏移量。这里下一条记录不是指插入顺序的下一条记录，而是主键值从小到大的顺序的下一条记录。Infimum的下一条记录就是主键值最小的用户记录，本页中主键值最大的用户记录的下一条记录就是Supremum。
  - 删除操作：delete_mask置位；删除记录的上一条指向删除记录的下一条；n_owned减一。无论怎么增删改查，InnoDB始终会维护一条记录的单链表，链表中的各个节点是按照主键值由小到大的顺序连接起来的。
  - 添加操作：如果将删除操作中的记录重新添加回来，直接复用原来被删除记录的存储空间。
  - 当数据页中存在多条被删除掉的记录时，这些记录的next_record属性将会把这些被删除掉的记录组成一个垃圾链表，以备之后重用这部分存储空间。

2.2.3 Infimum + Supremum（最小最大记录）

这两条记录不是我们定义的最大最小记录，并不存放在User Records部分，而是单独放在Infimum + Supremum部分。



2.3 第3部分：页目录和页面头部

2.3.1 Page Directory（页目录）

为什么需要页目录?
在页中，记录是以`单向链表`的形式进行存储的。单向链表的特点就是插入、删除非常方便，但是`检索效率不高`，最差的情况下需要遍历链表上的所有节点才能完成检索。因此在页结构中专门设计了页目录这个模块，`专门给记录做一个目录`，通过`二分查找法`的方式进行检索，提升效率。

页目录，二分法查找

1. 将所有的记录`分成几个组`，这些记录包括最小记录和最大记录，但不包括标记为“已删除”的记录。
2. 第 1 组，也就是最小记录所在的分组只有 1 个记录；最后一组，就是最大记录所在的分组，会有 1-8 条记录；其余的组记录数量在 4-8 条之间。
   这样做的好处是，除了第 1 组（最小记录所在组）以外，其余组的记录数会`尽量平分`。
3. 在每个组中最后一条记录的头信息中会存储该组一共有多少条记录，作为 n_owned 字段。
4. `页目录用来存储每组最后一条记录的地址偏移量`，这些地址偏移量会按照`先后顺序存储`起来，每组的地址偏移量也被称之为`槽（slot）`，每个槽相当于指针指向了不同组的最后一个记录。

分组步骤

- 初始情况下一个数据页里只有最小记录和最大记录两条记录，它们分属于两个分组。
- 之后每插入一条记录，都会从页目录中找到主键值比本记录的主键值大并且差值最小的槽，然后把该槽对应的记录的n_owned值加1，表示本组内又添加了一条记录，直到该组中的记录数等于8个。
- 在一个组中的记录数等于8个后再插入一条记录时，会将组中的记录拆分成两个组，一个组中4条记录，另一个5条记录。这个过程会在页目录中新增一个槽来记录这个新增分组中最大的那条记录的偏移量。

遍历一个组代价比较小，所以效率会高。

2.3.2 Page Header（页面头部）

占用固定的56个字节，专门存储各种状态信息。

名称	占用空间大小	描述
PAGE_N_DIR_SLOTS	2字节	在页目录中的槽数量
PAGE_HEAP_TOP	2字节	还未使用的空间最小地址，也就是说从该地址之后就是`Free Space`
PAGE_N_HEAP	2字节	本页中的记录的数量（包括最小和最大记录以及标记为删除的记录）
PAGE_FREE	2字节	第一个已经标记为删除的记录的记录地址（各个已删除的记录通过`next_record`也会组成一个单链表，这个单链表中的记录可以被重新利用）
PAGE_GARBAGE	2字节	已删除记录占用的字节数
PAGE_LAST_INSERT	2字节	最后插入记录的位置
PAGE_DIRECTION	2字节	记录插入的方向（若插入的比上一条记录的主键值大，则向右，否则向左，这里表示最后一条记录的插入方向）
PAGE_N_DIRECTION	2字节	一个方向连续插入的记录数量
PAGE_N_RECS	2字节	该页中记录的数量（不包括最小和最大记录以及被标记为删除的记录）
PAGE_MAX_TRX_ID	8字节	修改当前页的最大事务ID，该值仅在二级索引中定义
PAGE_LEVEL	2字节	当前页在B+树中所处的层级
PAGE_INDEX_ID	8字节	索引ID，表示当前页属于哪个索引
PAGE_BTR_SEG_LEAF	10字节	B+树叶子段的头部信息，仅在B+树的Root页定义
PAGE_BTR_SEG_TOP	10字节	B+树非叶子段的头部信息，仅在B+树的Root页定义

2.4 从数据页角度看B + 树如何查询

1.B+树是如何进行记录检索的?

如果通过B+树的索引查询行记录，首先是从B+树的根开始，逐层检索，直到找到叶子节点，也就是找到对应的数据页为止，将数据页加载到内存中，页目录中的槽(slot)采用`二分查找`的方式先找到一个粗略的记录分组然后再在分组中通过`链表遍历`的方式查找记录。

2.普通索引和唯一索引在查询效率上有什么不同?

唯一索引就是在普通索引上增加了约束性，也就是关键字唯一，找到了关键字就停止检索。而普通索引，可能会存在用户记录中的关键字相同的情况，根据页结构的原理，当我们读取一条记录的时候，不是单独将这条记录从磁盘中读出去，而是将这个记录所在的页加载到内存中进行读取。

InnoDB存储引擎的页大小为16KB，在一个页中可能存储着上千个记录，因此在普通索引的字段上进行查找也就是在内存中多几次“`判断下一条记录`”的操作，对于CPU来说，这些操作所消耗的时间是可以忽略不计的。所以对一个索引字段进行检索，采用普通索引还是唯一索引在检索效率上基本上没有差别。

3. InnoDB行格式(或记录格式)

我们平时的数据以行为单位来向表中插入数据，这些记录在磁盘上的存放方式也被称为行格式或者记录格式。

InnoDB存储引擎设计了4种不同类型的行格式

- Compact（紧密）（MySQL 5.1版本的默认行格式）
- Redundant（冗余）
- Dynamic（动态）
- Compressed（压缩）行格式

查看行格式

```
mysql> select @@innodb_default_row_format;
# 查询单张表行格式
mysql> show table status like 'departments' \G
```

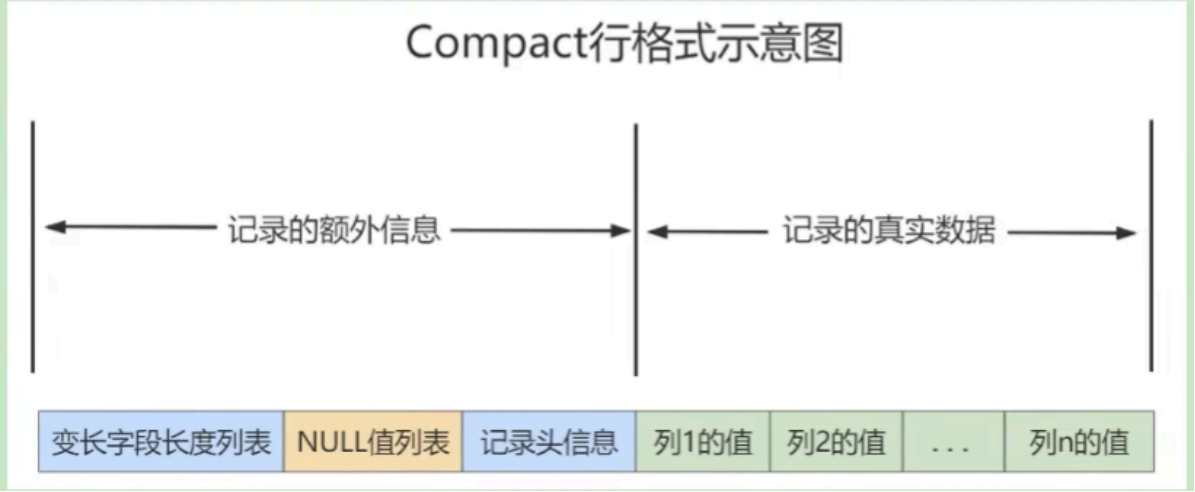
指定行格式的语法

```
CREATE TABLE 表名 (列的信息) ROW_FORMAT = 行格式名称;

ALTER TABLE 表名 ROW_FORMAT = 行格式名称;
```

3.1 Compact行格式

在MySQL 5.1版本中，默认设置为Compact行格式。一条完整的记录其实可以被分为记录的额外信息和记录的真实数据两大部分。



3.1.1 记录的额外信息

- 变长字段长度列表
  - 把所有变长字段（比如VARCHAR(M)、VARBINARY(M)、TEXT类型，BLOB类型等存储字节数不固定的类型）的真实数据占用的字节长度都存放在记录的开头部位，从而形成一个变长字段长度列表。
  - 注意：这里面存储的变长长度和字段顺序是反过来的。比如两个varchar字段在表结构的顺序是a(10)，b(15)。那么在变长字段长度列表中存储的长度顺序就是15，10，是反过来的。
  - 举个例子

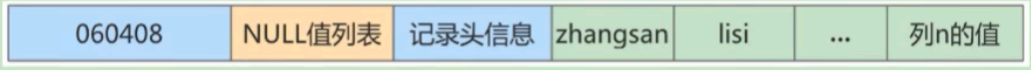
以record_test_table表中的第一条记录举例：因为record_test_table表的col1、col2、col4列都是VARCHAR(8)类型的，所以这三个列的值的长度都需要保存在记录开头处，注意record_test_table表中的各个列都使用的是ascii字符集（每个字符只需要1个字节来进行编码）。

列名	存储内容	内容长度（十进制表示）	内容长度（十六进制表示）
col1	'zhangsan'	8	0x08
col2	'lisi'	4	0x04
col4	'songhk'	6	0x06

又因为这些长度值需要按照列的逆序存放，所以最后变长字段长度列表的字节串用十六进制表示的效果就是（各个字节之间实际上没有空格，用空格隔开只是方便理解）：

06 04 08

把这个字节串组成的变长字段长度列表填入上边的示意图中的效果就是：



- NULL值列表
  - Compact行格式会把可以为NULL的列统一管理起来，存在一个标记为NULL值列表中。如果表中没有允许存储 NULL 的列，则 NULL值列表也不存在了。
  - **为什么定义NULL值列表?**
    - 如果不在记录内容中标注NULL则会造成混乱，如果使用特定符号标记NULL则会浪费空间，所以在行数据得头部开辟出一块空间专门用来记录该行数据哪些是非空数据，哪些是空数据.
  - 格式如下：
    1. 二进制位的值为1时，代表该列的值为NULL。
    2. 二进制位的值为0时，代表该列的值不为NULL。
  - 注意：同样顺序也是反过来存放的
  - 主键一定不为NULL，所以在NULL值列表中会跳过主键，也会跳过明确非空的键。
- 记录头信息（5字节）
  -

名称	大小（单位：bit）	描述
预留位1	1	没有使用
预留位2	1	没有使用

名称	大小（单位：bit)	描述
delete_mask	1	标记该记录是否被删除
mini_rec_mask	1	B+树的每层非叶子节点中的最小记录都会添加该标记
n_owned	4	表示当前记录拥有的记录数
heap_no	13	表示当前记录在记录堆的位置信息
record_type	3	表示当前记录的类型，0表示普通记录，1表示B+树非叶子节点记录，2表示最小记录，3表示最大记录
next_record	16	表示下一条记录的相对位置

- 前面已经说过了

3.1.2 记录的真实数据

- row_id
  - InnoDB机制中有聚簇索引，以主键建立的B+树叶子节点中存储了记录的全部信息，若未指定主键，则innoDB会指定一个隐藏主键，即row_id。
- transaction_id和roll_pointer后面讲。

3.3 Dynamic和Compressed行格式

3.3.1 行溢出

我们可以知道一个页的大小一般是16KB，也就是16384字节，而一个VARCHAR(M)类型的列就最多可以存储65533个字节，这样就可能出现一个页存放不了一条记录，这种现象称为`行溢出`

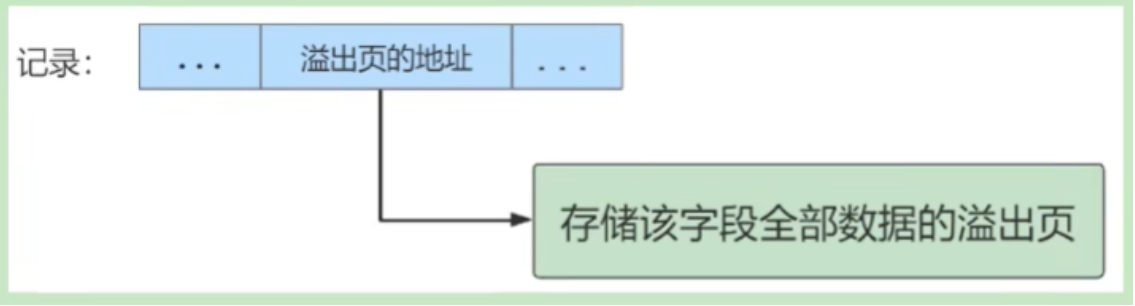
3.3.2 分页存储和页的扩展

在Compact和Reduntant行格式中，对于占用存储空间非常大的列，在记录的真实数据处只会存储该列的一部分数据，把剩余的数据分散存储在几个其他的页中进行`分页存储`。

然后记录的真实数据处用20个字节存储指向这些页的地址（当然这20个字节中还包括这些分散在其他页面中的数据的占用的字节数），从而可以找到剩余数据所在的页。这称为`页的扩展`。

3.3.3 不同行格式处理行溢出的区别

- Compressed和Dynamic两种记录格式对于存放在BLOB中的数据采用了完全的行溢出的方式。如图，在数据页中只存放20个字节的指针（溢出页的地址），实际的数据都存放在Off Page（溢出页）中。



- Compact和Redundant两种格式会在记录的真实数据处存储一部分数据（存放768个前缀字节）。

4. 区、段与碎片区

4.1 为什么要有区?

为了利用**磁盘的预读特性**。让区中存储连续的页，避免随机I/O。

查看4.n 扩展 理解mysql如何利用预读特性

`B+`树的每一层中的页都会形成一个双向链表，如果是以`页为单位`来分配存储空间的话，双向链表相邻的两个页之间的`物理位置`可能离得非常远。

范围查询只需要定位到最左边的记录和最右边的记录，然后沿着双向链表一直扫描就可以了，而如果链表中相邻的两个页物理位置离得非常远，就是所谓的`随机I/O`，随机I/O很慢，尽量让链表中相邻的页的物理位置也相邻。**这样利用了磁盘的预读特性。**

引入`区`的概念，一个区就是在物理位置上**连续**的`64个页`。因为InnoDB 中的页大小默认是16KB，所以一个区的大小是64*16KB=`1MB`。在表中`数据量大`的时候，为某个索引分配空间的时候就不再按照页为单位分配了，而是按照`区为单位`分配，甚至在表中的数据特别多的时候，可以一次性分配多个连续的区。虽然可能造成`一点点空间的浪费`（数据不足以填充满整个区），但是从性能角度看，可以消除很多的随机I/O，`功大于过`！

> 这里是连续的64个页， 但是具体的两个页之间还是用指针相连的。保证一大块区域连续。

4.2 为什么要有段?

为了对叶子节点和非叶子节点进行区别对待。

对于范围查询，其实是对B+树叶子节点中的记录进行顺序扫描，而如果不区分叶子节点和非叶子节点，统统把节点代表的页面放到申请到的区中的话，进行范围扫描的效果就大打折扣了。所以InnoDB对B+树的`叶子节点`和`非叶子节点`进行了区别对待，也就是说叶子节点有自己独有的区，非叶子节点也有自己独有的区。存放叶子节点的区的集合就算是一个`段(segment)`，存放非叶子节点的区的集合也算是一个段。也就是说一个索引会生成2个段，一个`叶子节点段`，一个`非叶子节点段`。

除了索引的叶子节点段和非叶子节点段之外，InnoDB中还有为存储一些特殊的数据而定义的段，比如回滚段。所以，常见的段有`数据段`、`索引段`、`回滚段`。数据段即为B+树的叶子节点，索引段即为B+树的非叶子节点。

在InnoDB存储引擎中，对段的管理都是由引擎自身所完成，DBA不能也没有必要对其进行控制。这从一定程度上简化了DBA对于段的管理。

段其实不对应表空间中某一个连续的物理区域，而是一个逻辑上的概念，由若干个零散的页面以及一些完整的区组成。

> 零散的页面，看碎片区

4.3 为什么要有碎片区?

为了避免数据量小的表浪费整片区的空间。

默认情况下，一个使用InnoDB存储引擎的表只有一个聚簇索引，一个索引会生成2个段，而段是以区为单位申请存储空间的，一个区默认占用1M (64*16Kb=1024Kb）存储空间，所以**默认情况下一个只存了几条记录的小表也需要2M的存储空间么?**以后每次添加一个索引都要多申请2M的存储空间么?这对于存储记录比较少的表简直是天大的浪费。这个问题的症结在于到现在为止我们介绍的区都是非常**纯粹的**，也就是一个区被整个分配给某一个段，或者说区中的所有页面都是为了存储同一个段的数据而存在的，即使段的数据填不满区中所有的页面，那余下的页面也不能挪作他用。

为了考虑以完整的区为单位分配给某个段对于`数据量较小`的表太浪费存储空间的这种情况，InnoDB提出了一个`碎片(fragment)区`的概念。在一个碎片区中，并不是所有的页都是为了存储同一个段的数据而存在的，而是碎片区中的页可以用于不同的目的，比如有些页用于段A，有些页用于段B，有些页甚至哪个段都不属于。`碎片区直属于表空间`，并不属于任何一个段。

所以此后为某个段分配存储空间的策略是这样的：

- 在刚开始向表中插入数据的时候，段是从某个碎片区以单个页面为单位来分配存储空间的
- 当某个段已经占用了`32个碎片区`页面之后，就会申请以完整的区为单位来分配存储空间。

所以现在段不能仅定义为是某些区的集合，更精确的应该是**某些零散的页面**以及**一些完整的区**的集合。

4.4 区的分类

区大体上可以分为4种类型:

- `空闲的区(FREE)`：现在还没有用到这个区中的任何页面。
- `有剩余空间的碎片区(FREE_FRAG)：` 表示碎片区中还有可用的页面。
- `没有剩余空间的碎片区(FULL_FRAG)`：表示碎片区中的所有页面都被使用，没有空闲页面。
- `附属于某个段的区(FSEG)：`每一个索引都可以分为叶子节点段和非叶子节点段。

处于`FREE`、`FREE_FRAG`以及`FULL_FRAG`这三种状态的区都是独立的，直属于表空间。而处于FSEG状态的区是附属于某个段的。

> 如果把表空间比作是一个集团军，段就相当于师，区就相当于团。一般的团都是隶属于某个师的，就像是处于`FSEG`的区全都隶属于某个段，而处于`FREE`、`FREE_FRAG`以及`FULL_FRAG`这三种状态的区却直接隶属于表空间，就像独立团直接听命于军部一样。

4.n 扩展

那么，计算机怎样才能判断一个数据接下来可能被用到?

> 时间局部性（Temporal Locality）
>
> 空间局部性（Spatial Locality）
>
> 顺序局部性（Order Locality）
>
> - 在典型程序中，除转移类指令外，大部分指令是顺序进行的。顺序执行和非顺序执行的比例大致是5:1。
> - 指令的顺序执行、数组的连续存放等是产生顺序局部性的原因。
> - 正在执行的某个指令以及还在排队等候处理的指令，大部分是按照顺序来执行的。

- 磁盘预读原理：磁盘读取的时候会顺带加载附近的数据到缓存，利用局部性
- 磁盘读取（计组的内容）
  - 磁盘存取，磁盘I/O涉及机械操作。磁盘是由大小相同且同轴的圆形盘片组成，磁盘可以转动(各个磁盘须同时转动)。磁盘的一侧有磁头支架，磁头支架固定了一组磁头，每个磁头负责存取一个磁盘的内容。磁头不动，磁盘转动，但磁臂可以前后动，用于读取不同磁道上的数据。磁道就是以盘片为中心划分出来的一系列同心环。磁道又划分为一个个小段，叫扇区，是磁盘的最小存储单元。
  - 磁盘读取时，系统将数据逻辑地址传给磁盘，磁盘的控制电路会解析出物理地址（哪个磁道，哪个扇区），于是磁头需要前后移动到相应的磁道——寻道，消耗的时间叫——寻道时间，磁盘旋转将对应的扇区转到磁头下（磁头找到对应磁道的对应扇区），消耗的时间叫——旋转时间，这一系列操作是非常耗时。

重点

预读和缺页

为了尽量减少I/O操作，计算机系统一般采取预读的方式，预读的长度一般为页（page）的整倍数。页是计算机管理存储器的逻辑块，硬件及操作系统往往将主存和磁盘存储区分割为连续的大小相等的块，每个存储块称为一页（在许多操作系统中，页得大小通常为4k），主存和磁盘以页为单位交换数据。当程序要读取的数据不在主存中时，会触发一个缺页异常，此时系统会向磁盘发出读盘信号，磁盘会找到数据的起始位置并向后`连续读取一页或几页载入内存中`，然后异常返回，程序继续运行。

计算机系统是分页读取和存储的，一般一页为4KB（8个扇区，每个扇区125B，8*125B=4KB），每次读取和存取的最小单元为一页，而`磁盘预读时通常会读取页的整倍数`。根据文章上述的【局部性原理】①当一个数据被用到时，其附近的数据也通常会马上被使用。②程序运行期间所需要的数据通常比较集中。由于磁盘顺序读取的效率很高（不需要寻道时间，只需很少的旋转时间），所以即使只需要读取一个字节，磁盘也会读取一页的数据。

至于磁盘分页，参考计算机操作系统的分页，分段存储管理——逻辑地址和物理地址被分为大小相同的页面，逻辑地址中叫页，物理地址中叫块。

为什么使用B-Tree/B+Tree

二叉查找树进化品种的红黑树等数据结构也可以用来实现索引，但是文件系统及数据库系统普遍采用B-Tree/B+Tree作为索引结构。

索引本身很大，一般存储在磁盘上。为了减少I/O次数，选择B树或B+树。（新建节点时直接申请一个页的空间，和物理空间的页对齐，读取一个node只需要一次I/O。

在树的深度上，红黑树等都不如B-Tree/B+Tree的低。

B树 VS B+Tree

B-Tree：如果一次检索需要访问4个节点，数据库系统设计者利用磁盘预读原理，把节点的大小设计为一个页，那读取一个节点只需要一次I/O操作，完成这次检索操作，最多需要3次I/O(根节点常驻内存)。数据记录越小，每个节点存放的数据就越多，树的高度也就越小，I/O操作就少了，检索效率也就上去了。

B+Tree：非叶子节点只存key，大大滴减少了非叶子节点的大小，那么每个节点就可以存放更多的记录，树更矮了，I/O操作更少了。所以B+Tree拥有更好的性能。

5. 表空间

表空间数据库由一个或多个表空间组成，表空间从管理上可以划分为

- `系统表空间` (System tablespace)
- `独立表空间` (File-per-table tablespace)
- `撤销表空间` (Undo Tablespace)
- `临时表空间` (Temporary Tablespace)

5.1 独立表空间

独立表空间，即每张表有一个独立的表空间，也就是数据和索引信息都会保存在自己的表空间中。独立的表空间(即：单表)可以在不同的数据库之间进行`迁移`。

空间可以回收(DROPTABLE操作可自动回收表空间；其他情况，表空间不能自己回收)。如果对于统计分析或是日志表，删除大量数据后可以通过：`alter table TableName engine = innodb`；回收不用的空间。对于使用独立表空间的表，不管怎么删除，表空间的碎片不会太严重的影响性能，而且还有机会处理。

独立表空间结构

独立表空间由段、区、页组成。前面已经讲解过了。

真实表空间对应的文件大小

我们到数据目录里看，会发现一个新建的表对应的`.ibd`文件只占用了`96K`，才6个页面大小(MySQL5.7中)，这是因为一开始表空间占用的空间很小，因为表里边都没有数据。不过别忘了这些.ibd文件是`自扩展的`，随着表中数据的增多，表空间对应的文件也逐渐增大。

查看InnoDB的表空间类型语句

```
# 查看是否独立表空间
mysql> show variables like 'innodb_file_per_table';
```

5.2 系统表空间

系统表空间的结构和独立表空间基本类似，只不过由于整个MySQL进程只有一个系统表空间，在系统表空间中会额外记录一些有关整个系统信息的页面，这部分是独立表空间中没有的。

InnoDB数据字典

每当我们向一个表中插入一条记录的时候，`MySQL校验过程`如下:

先要校验一下插入语句对应的表存不存在，插入的列和表中的列是否符合，如果语法没有问题的话，还需要知道该表的聚簇索引和所有二级索引对应的根页面是哪个表空间的哪个页面，然后把记录插入对应索引的B+树中。所以说，MySQL除了保存着我们插入的用户数据之外，还需要保存许多额外的信息，比方说:

- 某个表属于哪个表空间，表里边有多少列
- 表对应的每一个列的类型是什么
- 该表有多少索引，每个索引对应哪几个字段，该索引对应的根页面在哪个表空间的哪个页面
- 该表有哪些外键，外键对应哪个表的哪些列
- 某个表空间对应文件系统上文件路径是什么
- ...

上述这些数据并不是我们使用`INSERT`语句插入的用户数据，实际上是为了更好的管理我们这些用户数据而不得已引入的一些额外数据，这些数据也称为`元数据`。InnoDB存储引擎特意定义了一些列的`内部系统表`(internalsystem table)来记录这些这些元数据:

表名	描述
`SYS_TABLES`	整个InnoDB存储引擎中所有的表的信息
`SYS_COLUMNS`	整个InnoDB存储引擎中所有的列的信息
`SYS_INDEXES`	整个InnoDB存储引擎中所有的索引的信息
`SYS_FIELDS`	整个InnoDB存储引擎中所有的索引对应的列的信息
SYS_FOREIGN	整个InnoDB存储引擎中所有的外键的信息
SYS_FOREIGN_COLS	整个InnoDB存储引擎中所有的外键对应列的信息
SYS_TABLESPACES	整个InnoDB存储引擎中所有的表空间信息

表名	描述
SYS_DATAFILES	整个InnoDB存储引擎中所有的表空间对应文件系统的文件路
SYS_VIRTUAL	整个InnoDB存储引擎中所有的虚拟生成列的信息

这些系统表也被称为`数据字典`，它们都是以`B+`树的形式保存在系统表空间的某些页面中，其中`SYS_TABLES`、`SYS_COLUNNS`、`SYS_INDEXES`、`SYS_FIELDS`这四个表尤其重要，称之为基本系统表(basic system tables)

注意:用户是`不能直接访问`InnoDB的这些内部系统表，除非你直接去解析系统表空间对应文件系统上的文件。不过考虑到查看这些表的内容可能有助于大家分析问题，所以在系统数据库`information_schema`中提供了一些以innodb_sys开头的表。

在`information_schema`数据库中的这些以`INNODB_SYS`开头的表并不是真正的内部系统表(内部系统表就是我们上边以SYS开头的那些表)，而是在存储引擎启动时读取这些以`SYS`开头的系统表，然后填充到这些以
`INNODB_SYS`开头的表中。以`INNODB_SYS`开头的表和以`SYS`开头的表中的字段并不完全一样，但供大家参考已经足矣。

附录: 数据页加载的三种方式

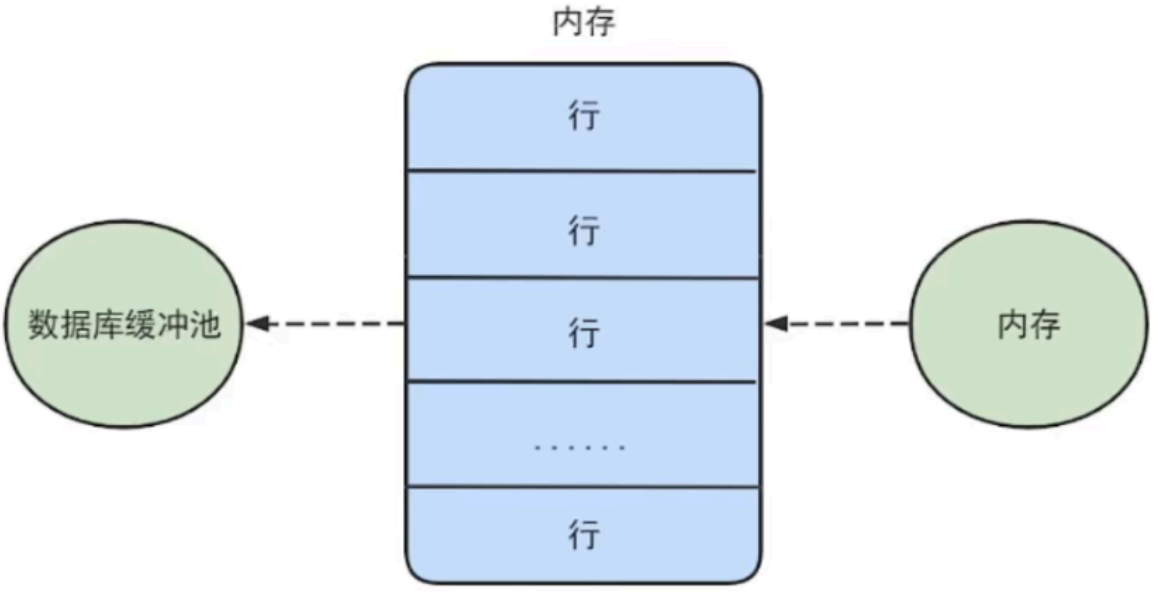
InnoDB从磁盘中读取数据的`最小单位`是数据页。而你想得到的id = xoxx的数据，就是这个数据页众多行中的一行。

对于MySQL存放的数据，逻辑概念上我们称之为表，在磁盘等物理层面而言是`按数据页`形式进行存放的，当其加载到MysQL中我们称之为`缓存页`。

如果缓冲池中没有该页数据，那么缓冲池有以下三种读取数据的方式,每种方式的读取效率都是不同的:

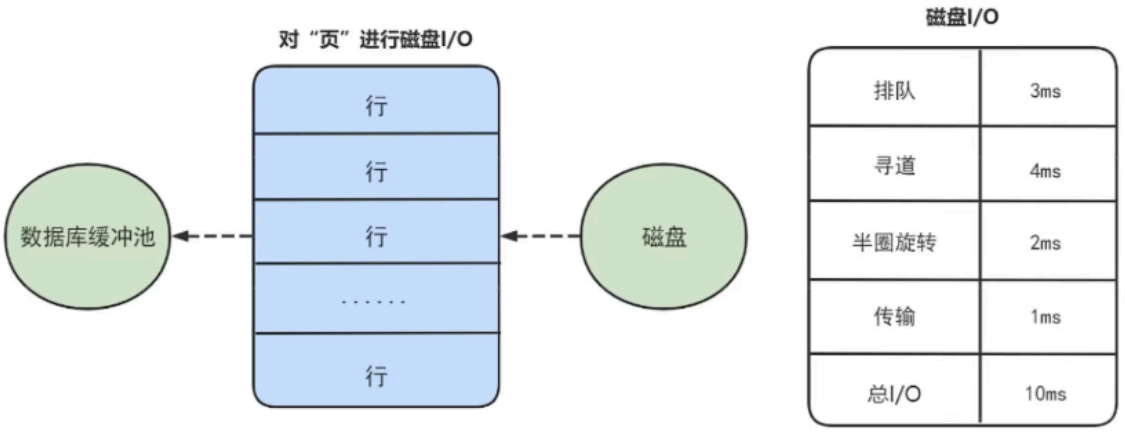
1. 内存读取

如果该数据存在于内存中，基本上执行时间在1ms左右，效率还是很高的。



2. 随机读取

如果数据没有在内存中，就需要在磁盘上对该页进行查找，整体时间预估在`10ms`左右，这10ms 中有6ms是磁盘的实际繁忙时间(包括了`寻道和半圈旋转时间`），有3ms是对可能发生的排队时间的估计值，另外还有1ms的传输时间，将页从磁盘服务器缓冲区传输到数据库缓冲区中。这10ms 看起来很快，但实际上对于数据库来说消耗的时间已经非常长了，因为这还只是一个页的读取时间。



3. 顺序读取

顺序读取其实是一种批量读取的方式，因为我们请求的`数据在磁盘上往往都是相邻存储的`，顺序读取可以帮我们批量读取页面，这样的话，一次性加载到缓冲池中就不需要再对其他页面单独进行磁盘I/O操作了。如果一个磁盘的吞吐量是40MB/S，那么对于一个16KB大小的页来说，一次可以顺序读取2560(40MB/16KB)个页，相当于一个页的读取时间为0.4ms。采用批量读取的方式，即使是从磁盘上进行读取，效率也比从内存中只单独读取一个页的效率要高。