

Visitors

CN

177,205

US

15,556

HK

5,780

JP

3,721

TW

3,008

SG

1,356

CA

816

FR

805

GB

775

AU

695

DE

670

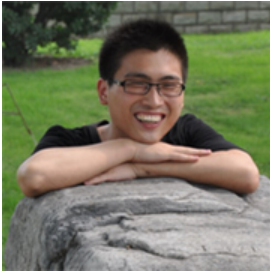
KR

336

Pageviews: 377,090

Flags Collected: 83

FLAG counter



个人经历

2015 至今 阿里巴巴

2013-2015 美团

2010-2013 中科院 (硕士)

2006-2010 浙大 (本科)

阿里巴巴RDC长期招聘Java研发工程师, 有意者站内联系!

昵称: zhanlijun

园龄: 4年10个月

粉丝: 664

关注: 5

+加关注

最新随笔

1. 一个复杂系统的拆分改造实践

2. mysql死锁问题分析

3. 近期code review几处小问题集锦

4. 你应该知道的RPC原理

5. 如何健壮你的后端服务?

6. 如何用消息系统避免分布式事务?

7. 一个故事讲清楚NIO

8. 地图匹配实践

9. 利用模拟退火提高Kmeans的聚类精度

10. 空间插值文献阅读 (Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall)

位图索引:原理 (BitMap index)

<http://www.cnblogs.com/LBSer/p/3322630.html>

位图 (BitMap) 索引

前段时间听同事分享, 偶尔讲起Oracle数据库的位图索引, 顿时大感兴趣。说来惭愧, 在这之前对位图索引一无所知, 因此趁此机会写篇博文介绍下位图索引。

1. 案例

有张表名为table的表, 由三列组成, 分别是姓名、性别和婚姻状况, 其中性别只有男和女两项, 婚姻状况由已婚、未婚、离婚这三项, 该表共有100w个记录。现在有这样的查询: `select * from table where Gender='男' and Marital='未婚';`

姓名(Name)	性别(Gender)	婚姻状况(Marital)
张三	男	已婚
李四	女	已婚
王五	男	未婚
赵六	女	离婚
孙七	女	未婚
...

1) 不使用索引

不使用索引时, 数据库只能一行行扫描所有记录, 然后判断该记录是否满足查询条件。

2) B树索引

对于性别, 可取值的范围只有'男','女', 并且男和女可能各站该表的50%的数据, 这时添加B树索引还是需要取出一半的数据, 因此完全没有必要。相反, 如果某个字段的取值范围很广, 几乎没有重复, 比如身份证号, 此时使用B树索引较为合适。事实上, 当取出的行数据占用表中大部分的数据时, 即使添加了B树索引, 数据库如oracle、mysql也不会使用B树索引, 很有可能还是一行行全部扫描。

2. 位图索引出马

如果用户查询的列的基数非常的小, 即仅有的几个固定值, 如性别、婚姻状况、行政区等等。要为这些基数值比较小的列建索引, 就需要建立位图索引。

对于性别这个列, 位图索引形成两个向量, 男向量为10100..., 向量的每一位表示该行是否是男, 如果是则位1, 否则0, 同理, 女向量位01011。

RowId	1	2	3	4	5	
男	1	0	1	0	0	
女	0	1	0	1	1	

随笔分类(57)

java(3)

LBS(10)

paper阅读笔记(2)

大数据(6)

定位原理/算法(3)

发表的SCI/SSCI(4)

服务治理(4)

空间索引原理(7)

数据库(5)

推荐相关(1)

线上问题定位及解决(2)

消息系统(2)

信息检索算法/实践(6)

应用服务器(2)

积分与排名

积分 - 115075

排名 - 2612

最新评论

1. Re:如何设计实现一个地址反解析服务？

如果仅仅是为了将用户坐标解析到道路级别的话，也未必需要用栅格。对于任意一条道路，根据历史记录，可以得到定位于这条道路的所有点，根据这堆点可以得到一个外包多边形，以后所有落在这个多边形内的点都可以认为是.....

--张可纯biubiu

2. Re:GeoHash核心原理解析

lucene里面使用了geohash，但是计算距离的时候貌似还是用经纬度计算距离，那使用geohash还有什么意义呢？

--casterQL

阅读排行榜

1. GeoHash核心原理解析(43980)

对于婚姻状况这一列，位图索引生成三个向量，已婚为11000...，未婚为00100...，离婚为00010....。

RowId	1	2	3	4	5	
已婚	1	1	0	0	0	
未婚	0	0	1	0	1	
离婚	0	0	0	1	0	

当我们使用查询语句“select * from table where Gender=‘男’ and Marital=“未婚”；”的时候首先取出男向量10100...，然后取出未婚向量00100...，将两个向量做and操作，这时生成新向量00100...，可以发现第三位为1，表示该表的第三行数据就是我们需要查询的结果。

RowId	1	2	3	4
男	1	0	1	0
and				
未婚	0	0	1	0
结果	0	0	1	0

3.位图索引的适用条件

上面讲了，位图索引适合只有几个固定值的列，如性别、婚姻状况、行政区等等，而身份证号这种类型不适合用位图索引。

此外，位图索引适合静态数据，而不适合索引频繁更新的列。举个例子，有这样一个字段busy，记录各个机器的繁忙与否，当机器忙碌时，busy为1，当机器不忙碌时，busy为0。

这个时候有人会说使用位图索引，因为busy只有两个值。好，我们使用位图索引索引busy字段！假设用户A使用update更新某个机器的busy值，比如update table set table.busy=1 where rowid=100；，但还没有commit，而用户B也使用update更新另一个机器的busy值，update table set table.busy=1 where rowid=12；这个时候用户B怎么也更新不了，需要等待用户A commit。

原因：用户A更新了某个机器的busy值为1，会导致所有busy为1的机器的位图向量发生改变，因此数据库会将busy = 1的所有行锁定，只有commit之后才解锁。

转载请标明源地址：http://www.cnblogs.com/LBSer

分类: 空间索引原理

标签: BitMap 位图索引

好文要顶

关注我

收藏该文

zhanlijun

关注 - 5

粉丝 - 664

+加关注

« 上一篇：GeoHash核心原理解析

» 下一篇：NoSQL之Cassandra

160

posted @ 2013-09-15 15:52 zhanlijun 阅读(21132) 评论(20) 编辑 收藏

评论列表

#1楼 2013-09-15 17:44 大圆那些事

通俗易懂~

支持(0) 反对(0)

2. 你应该知道的RPC原理(30598)
3. 如何用消息系统避免分布式事务？(23677)
4. mysql死锁问题分析(22275)
5. 位图索引:原理 (BitMap index) (21132)

评论排行榜
1. 地图匹配实践(82)
2. 如何用消息系统避免分布式事务？(42)
3. 你应该知道的RPC原理(23)
4. GeoHash核心原理解析(22)
5. 地理围栏算法解析 (Geo-fencing) (20)

#2楼 2013-09-16 09:01 阿春阿晓	好文~	支持(0) 反对(0)
#3楼 2013-09-16 09:16 自由地飞翔	学习了这两个索引，还是比较有用的。	支持(0) 反对(0)
#4楼[楼主] 2013-09-16 10:07 zhanlijun	@ 蓝色天晶 一起学习！	支持(0) 反对(0)
#5楼[楼主] 2013-09-16 10:07 zhanlijun	@ 大圆那些事 thks	支持(0) 反对(0)
#6楼[楼主] 2013-09-16 10:08 zhanlijun	@ 阿春阿晓 感谢！	支持(0) 反对(0)
#7楼 2013-09-16 16:46 吴煜	恩，位图索引的内容和适用范围介绍的很详细，但是位图所以也有一个问题，比如需要搜索是select * from table where Gender='男' or Marital="未婚"的情况~	支持(0) 反对(0)
#8楼[楼主] 2013-09-16 19:22 zhanlijun	@ 吴煜 这样就得到10101...这样的向量，表明第1、3、5行是我们想要的结果，效果也很好啊	支持(1) 反对(0)
#9楼 2013-09-16 21:15 吴煜	对的，下午没有仔细比对一下，or运算用位图索引是能得到想得到的结果的，赞！	支持(1) 反对(0)
#10楼 2014-03-18 13:20 刀尖红叶	好文，通俗易懂	支持(0) 反对(0)
#11楼 2014-09-07 14:38 糖拌咸鱼	Bitmap 另一个关键特性的支持比较高效的压缩算法，可以减少storage。	支持(0) 反对(0)
#12楼[楼主] 2014-09-07 17:06 zhanlijun	@ 糖拌咸鱼 对的，位图本身也是一种节省空间的方法	支持(0) 反对(0)
#13楼 2014-10-11 10:35 dyc0113	有个问题需要问楼主： 如果我的语句是 select * from table where Gender='男'	

性别建立了位图索引，当我需要搜索性别为男的记录时，我需要把'男'对应的位图向量取出来，那么取出位图向量时，我仍然需要遍历整个位图向量来获取到底哪一位是 1，位图向量的长度跟表中记录的长度是一样的，此时效率岂不是还是很低吗？ 这种情况下，建立位图向量和不建立位图向量的区别只是在于磁盘遍历和内存遍历的速度问题吗？

支持(1) 反对(0)

#14楼[楼主] 2014-10-13 11:13 zhanlijun

@ dyc0113

你说的是对的，位图向量要拿到内存里，不过由于位图向量具有数据压缩功能，其对内存的消耗会很小

支持(0) 反对(0)

#15楼 2015-04-12 10:43 计算机的潜意识

@ dyc0113

简单说来，位图是用字节8个bit中的每个bit来代表0和1的，这样占用存储就非常小。以这张表为例，假设有100万的数据，那么代表“男性”的位图的大小就是100万Bit。注意，这只是Bit，换算成字节的话，就是12.5万Byte，也就是125KB的大小，这个大小是绝对足够全部装入内存的。**位图的关键是缩小了存储空间，以使得内存遍历成为可能。**

即便我有足够大的机器，能够把所有记录放入内存里，对bit位的遍历(判断0和1)和遍历一条记录(比较两个值)的速度也有非常大的差距。前者有直接的机器指令支持，后者则需要多条指令才能完成，在机器指令集方面就拉开了差距。同时在进行多条件判断时，位运算也有直接的机器指令，还支持指令集并行优化，因此综合起来，位图索引的效率就非常高了。

支持(4) 反对(0)

#16楼 2015-06-10 16:03 sjr1988

貌似只能使用两种状态的情况，如果是3种或者是4中状态的情况，怎么建索引？

支持(0) 反对(0)

#17楼[楼主] 2015-06-22 08:38 zhanlijun

@ sjr1988

当然可以支持多种状态了。现在仅仅是用了一个bit，一个bit只有0和1，因此只能表达两种状态；如果我使用两个bit，那就可以表达4种状态了，以此类推。当然表达的状态越多（使用的bit越多），消耗的存储也越大

支持(0) 反对(0)

#18楼 2016-11-14 10:19 GeorgeChong

第一，赞；第二，ID 这种类型不适合用位图索引，这不对，位图索引只是不合适频繁删除修改的应用场景，如果“势值”过大，可以采用分桶的方式，解决该问题；第三，bitmap index 更适用于数据库或者科学数据研究，最佳场景是，数据量大，又不频繁修改。

支持(0) 反对(0)

#19楼 2017-05-26 15:53 夜之悲哀

通俗易懂

支持(0) 反对(0)

#20楼 2017-07-31 15:51 shamyang

讲的非常好

支持(0) 反对(0)

刷新评论 刷新页面 返回顶部

注册用户登录后才能发表评论，请 [登录](#) 或 [注册](#)，访问网站首页。

【推荐】超50万VC++源码：大型组态工控、电力仿真CAD与GIS源码库！

【缅怀】传奇谢幕，回顾霍金76载传奇人生

【推荐】业界最快速.NET数据可视化图表组件

【腾讯云】买域名送解析+SSL证书+建站

【活动】2050 科技公益大会 - 年青人因科技而团聚



- 阿里云 新购满返 ¥6000 封顶

- 写给自学者的入门指南
- 和程序员谈恋爱
- 学会学习
- 优秀技术人的管理陷阱
- 作为一个程序员，数学对你到底有多重要

» [更多知识库文章...](#)