

列存储的压缩算法【转】

本次分享内容由三个部分组成:

- 列存储数据组织实现
- 智能索引的实现
- 自适应压缩算法

目前数据库主流数据组织技术分为**数据按行存储**和**数据按列存储**，达梦数据库表数据的存储方式同时支持行存储和列存储。行存储是以记录为单位进行存储的，数据页面中存储的是完整的若干条记录；列存储是以列为单位进行存储的，每一个列的所有行数据都存储在一起，一个段只存储一个列的数据，而且一个指定的页面中存储的都是某一个列的连续数据。

列存表的存储方式有以下几个优点:

同一个列的数据都是连续存储的，可以加快某一个列的数据查询速度（相对行存减小不必要IO）；



连续存储的列数据，具有更大的压缩单元和数据相似性，可以获得远优于行存储的压缩效率；

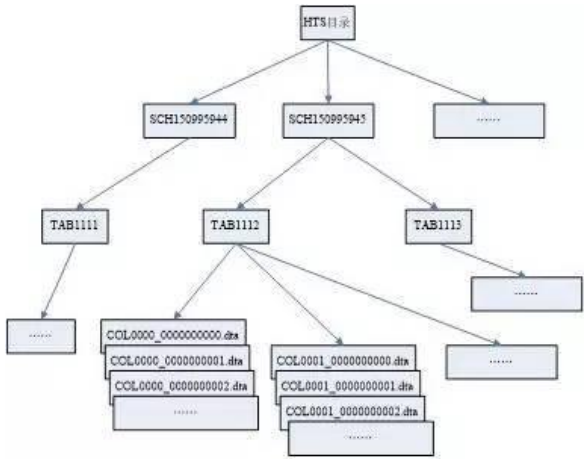
条件扫描借助数据区的统计信息进行精确过滤，可以进一步减少IO，提高扫描效率（智能索引）。

下面我们主要介绍达梦数据库的列存储的技术特点及具体实现，并详细介绍列存储的数据组织方式、智能索引的实现、自适应压缩算法设计。

1

列存储数据组织实现

达梦的列存储表又叫HUGE表，是建立在HTS表空间上的（全称HUGE TABLESPACE）。这个表空间与普通的表空间不同。普通的表空间，数据是通过段、簇、页来管理的，并且以固定大小（4K、8K、16K、32K）的页面为管理单位；而HTS相当于一个简单的文件系统，创建一个HTS，其实就是创建一个空的目录。在创建一个HFS表之后，数据库会在指定的HTS表空间目录下创建一系列的目录及文件，文件系统结构如下图所示：



从上图可以看出，在HTS目录下成功创建HFS表，系统内部需要经过以下步骤： 首先，在HTS目录下创建这个表对应的模式目录。目录名为“SCH+长度为9的ID号”组成的字符串。

其次，在模式目录下创建对应的表目录。表目录也是同样的道理，表目录名为“TAB+长度为4的ID号”组成的字符串。表目录中存放的是这个表中所有的文件。

再次，在新创建表时，每一个列对应的一个以dta为后缀的文件，文件大小可以在建表时指定，默认为64M，文件名为“COL+长度为4的列号+_+长度为10的文件号”。例如，在上图中，0000表示第1列，0001表示第2列.....；0000000000表示第1个文件，0000000001表示第2个文件.....最初一个列只有一个文件即可，当随着数据量不断增长，一个文件已放不下之后，系统会自动创建新的文件来存储不断增长的数据。

对于一个文件，其内部存储是按照区来管理的，区是文件内部数据管理的最小单位，也是唯一的单位（类似于行存的页）。一个区中，可以存储单列数据的行数是通过创建表时指定的，一经指定，在这个表的生命过程就不能再修改。

一个区内，存储的仅仅是数据而已，对于ABLE属性的列，存储的还有相应的标志。每一个区的开始位置及长度在文件内都是4K对齐的。对于一个HFS表，相应的还配备一个辅助表来管理其数据。

因为在上面介绍的文件中只存储了数据，辅助表用来管理以及辅助系统用户操作这些数据，该辅助表是在创建HFS表时系统自动创建的。辅助表中每一条记录对应文件中的一个数据区，辅助表包括下面15列：

”

1. COLID：表示当前这条记录对应的区所在的列的列ID号；
2. SEC_ID：表示当前这个记录对应的区的区ID号，每一个区都有一个ID号，并且唯一；
3. FILE_ID：表示这个区的数据所在的文件号；
4. OFFSET：表示这个区的数据在文件中的偏移位置，4K对齐；
5. COUNT：表示这个区中存储的数据总数（有可能包括被删除的数据）；
6. ACOUNT：表示这个区中存储的实际数据行数；
7. N_LEN：表示这个区中存储的数据在文件中的长度，4K对齐的；
8. N_：表示这个区中的数据中包括的值的行数；
9. N_DIST：表示这个区中所有数据互不相同的行数；
10. MAX_VAL：表示这个区中的最大值，精确值；
11. MIN_VAL：表示这个区中的最小值，精确值；
12. SUM_VAL：表示这个区中所有值的和，精确值；
13. CPR_FLAG：表示这个区是否压缩；
14. ENC_FLAG：表示这个区是否加密；
15. CHKSUM：用来校验的，该功能暂未启用。

”

前面7列是用来控制数据存取的，根据这些信息就可以知道这个区的具体存储位置、长度及基本信息。后面8列都是用来对这个区进行统计分析的。其中，COLID和SEC_ID的组合键为辅助表的聚簇关键字。

实际上数据库对列存数据的查找过程就是通过对辅助表信息的检索，利用辅助信息操纵HTS目录下文件的过程。

2

智能索引的实现

- 2014年2月 (1)
- 2014年1月 (51)
- 2013年12月 (1)
- 2013年10月 (1)
- 2013年2月 (1)
- 2012年10月 (7)
- 2012年9月 (24)

最新评论

- 1. Re:HashMap 和 ConcurrentHashMap, Java1.8版本
关于CounterCell的问题，可以看下CounterCell的注释。/** * A padded cell for distributing counts. Adapted from L.....
--superfl
- 2. Re:广告引擎解析
楼主也在从事广告行业？我也是，你现在的在哪家公司工作呢？
--lizongwu
- 3. Re:codis学习
你好，很高兴能看到你的这篇文章，最近在学习这个，我按照豌豆荚的安装步骤最后服务是起来了，但是代理服务名字有问题，dashboard页面上显示localhost:19000而不是ip:19000,用ja.....
--多宝鱼
- 4. Re:搭建packagist私服和composer
请问博主[Symfony\Component\DependancyInjection\Exception\InvalidArgumentException]
--血色残阳_1987
- 5. Re:Xapian简明教程(未完成)
这篇文章太赞了.可惜未完成.
请问博主,此文的原文在哪里呢?
--sirlipeng

阅读排行榜

- 1. 线程池几种配置参数的理解(10036)
- 2. Xapian简明教程(未完成)(4144)
- 3. 给定经纬度定位某个城市(3676)
- 4. JVM线程状态, park, wait, sleep, interrupt, yeild 对比(3462)
- 5. codis学习(1602)

评论排行榜

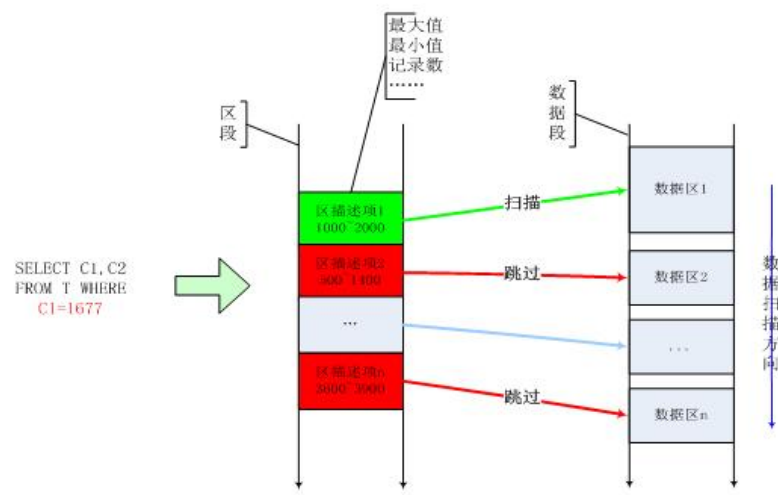
- 1. Xapian简明教程(未完成)(1)
- 2. 广告引擎解析(1)
- 3. HashMap 和 ConcurrentHashMap, Java1.8版本(1)
- 4. codis学习(1)
- 5. 搭建packagist私服和composer(1)

推荐排行榜

- 1. 一致性哈希节点解决雪崩问题(1)
- 2. 广告引擎解析(1)
- 3. mysql事务隔离级别(1)
- 4. flume-ng 自定义sink消费flume source(1)

列存表做条件扫描可以借助数据区的统计信息进行精确过滤，可以进一步减少IO，提高扫描效率。数据区的统计信息是数据库系统自动维护的，在一定程度上可以替代BTree索引，所以被一些厂商称为“智能索引”。

下图为区最大值、最小值信息起到过滤作用的示意图：



除了最大值、最小值起到跳区扫描减少IO的作用，区和值、区内有效行数等其它信息在一些统计分析场景甚至可以完全避免数据区的扫描。如MAX，COUNT，AVG等一些函数可以直接通过查询辅助表获得。

3

自适应压缩算法

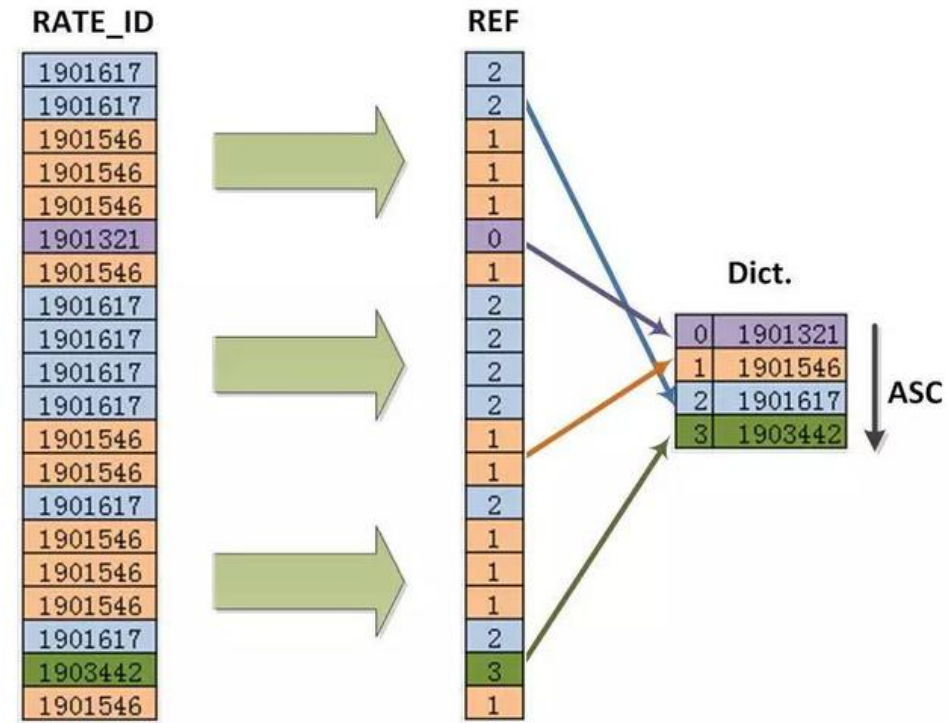
连续存储的列数据，具有更大的压缩单元和数据相似性，可以获得远优于行存储的压缩效率。根据列存储的特性，达梦提供一套自适应的压缩算法，可以根据数据特性以区为单位自动选择相应的算法最大化压缩率和性能。

下面详细介绍4种编码策略：

(1)字典编码

利用数据类型的一致性，将相同的值提取出来生成符号表，每个列值则直接存储该值映射成的符号表值id即可。

字典编码的示例如下：



(2)常量编码

当区内的数据大部分的数据相同，只有少数不同时，可以采用常量编码。该编码将区内数据出现最多的一个值作为常量值，其他值作为异常值。异常值使用<行号+值>的方式存储。

OPERATION_TYPE_ID

1	TTGW0001
2	TTGW0001
3	TTGW0001
4	TTGW0001
5	TTGW0001
6	TTGW0001
7	TTGW0001
8	TTGW0001
9	TTGW0001
10	TTGW0001
11	TTGW0003
12	TTGW0001
13	TTGW0001
14	TTGW0001
15	TTGW0001
16	TTGW0001
17	TTGW0001
18	TTGW0001
19	TTGW0001
20	TTGW0010

Default Value

TTGW0001

Exception Table

11	TTGW0003
20	TTGW0010

(3)RLE编码

当区内的数据存在大量的相同值，每个不同值的个数比较均匀，且连续出现时，可以使用RLE编码。

CALLED_NUMBER_COMPANY_ID

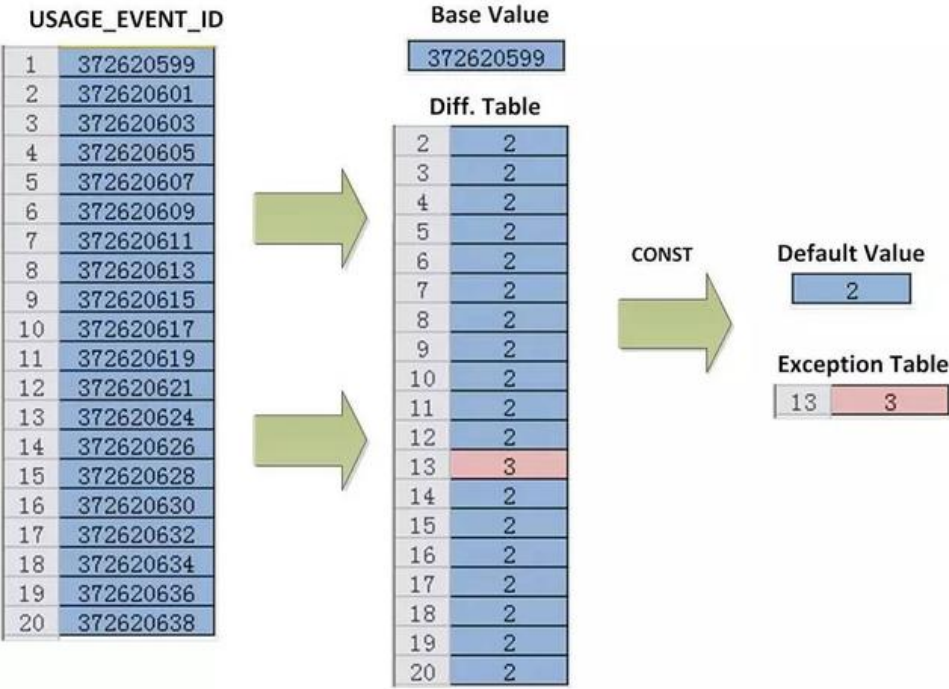
1	9001
2	9001
3	9001
4	9001
5	9001
6	9004
7	9001
8	9002
9	9002
10	9002
11	9002
12	9002
13	9003
14	9007
15	9007
16	9007
17	9007
18	9007
19	9007
20	9007

RLE Ref Table

1	9001
6	9004
7	9001
8	9002
13	9003
14	9007

(4)序列编码

当区内的数据差值成等差数列，或者存在一定的代数关系，则可以使用序列编码。



(5)四种编码使用策略

字典编码、常量编码和RLE编码都是基于重复值较多的情况下使用的，如果不同值中，某个值出现多次，其他的值都只出现一次，则可以使用常量编码；如果这些不同值都连续出现多次，且出现的次数相当，则使用RLE编码；如果这些不同值出现的次数相当，但都不连续，则使用字典编码。从上可知，这三种编码方式都是基于不同值的分布特征来确定使用哪种编码，也就是说，只会使用三种中的一种编码方式，不会嵌套使用。

序列编码基于前后两个数据的差值大部分都相同的一种编码方式，当差值计算出来之后，可以再分析其数据特征确定再使用前面三种编码方式。

综上所述，这四种编码策略如下：

“

1. 如果该列为自增列或序列，则直接使用序列编码；
2. 获得区数据统计信息：不同值个数n_dist、每个不同值个数、不同值数据指针、每个值连续出现的次数、整型数的最大值；
3. 根据获得的区统计信息，确定使用常量编码、RLE编码还是字典编码；这三种编码的使用顺序为：优先使用常量编码，其次使用RLE编码，最后才使用字典编码。
4. 如果编码后的总长度超过了原始长度，则不编码，直接返回；

”

除了采用编码进行压缩外还可以对区采用QUICKLZ或1-9级LZ算法进一步压缩，两者可以同时存在，也就是对编码后的存储结构进一步二进制压缩。

通过以上压缩算法，通常可以将数据压缩几倍，如果运气好（数据有序、重复值很多等）压缩率可达数百倍。目前CPU性能发展很快，而存储性能提升有限，特别是大数据处理领域不大可能大规模使用SSD。解压引起的性能损失远远小于磁盘IO等待的开销。

Q&A

Q1: 智能索引针对的数据类型和长度有限制吗？

A1: “智能索引”不能用于大字段类型如BLOB,CLOB，其它达梦支持的基本数据类型都可以利用“智能索引”。

Q2: 字典压缩算法的字典有大小限制吗？会根据列中数据量变化吗？字典建立在列上还是区上呢？

A2: 区是列的子集，如果数据库采用字典编码去压缩一个区，则这个区维护一个字典，字典会根据数据量变化。比如这个区有1000个不同值，则字典就有1000个值。有1万个不同值，则字典就有1万个值。字典编码有大小限制，如果系统判断采用字典压缩算法后，压缩率达不到阈值会放弃采用该编码压缩。

Q3: 首先，在HTS目录下创建这个表对应的模式目录。目录名为“SCH+长度为9的ID号”组成的字符串。9位的ID号，是指表的tabid么？

A3: SCH+9位ID，那个ID是模式的编码号。

Q4: 如果字段值为20000000000000000000xxxxxxx，这种数据类型，智能索引是否还会生效？

A4: 会生效，只有数据很乱的时候才不生效。比如所有的区最大值最小值都很相近，比如都是1到10000000000，则无法利用智能索引。

Q5: 数据字段中若有大量的空值，请问这种压缩策略能不能有很好的压缩率？如何实现？


- A5:** 在达梦数据库里也是一个特殊字符，在列存里同其它数据处理。也是根据数据的实际情况选择编码方式压缩。一般大量同一个字符，压缩率一般都很好。
- Q6:** 请问下收费规则如何？支持数据仓库吗？对经典的多对多关系怎么拆分？
- A6:** 数据库的收费规则得咨询销售，不过肯定比Oracle便宜。数据仓库可以基于我们构建，你说的多对多关系拆分可以线下讨论。
- Q7:** 和Oracle兼容性如何？
- A7:** 达梦在DML类语句上，100%兼容Oracle语法。PL/SQL大部分都兼容。如果DML语句发现有不兼容的地方告诉我们，我们立马改了。PL/SQL不兼容的看情况，对于常用的语法和Oracle是兼容的，很冷门的包如果没有项目推动，一般不会主动去兼容。
- Q8:** 达梦是基于开源数据库改的么？
- A8:** 完全自主开发、自主原创，源代码都是我们自己写的。

作者介绍 李鹏

- 现任职于达梦数据库，从事数据库测试和技术职称工作；
- 擅长数据库性能瓶颈分析定位及调优，主要负责数据库产品性能测试及重点项目竞争性测试；
- 曾参与国电电网、南方电网、国家工商总局等项目的POC测试，积累了丰富的数据库测试经验。

<http://www.58maisui.com/2016/06/17/a-218/>





[23lalala](#)
关注 - 1
粉丝 - 4

0

0

[+加关注](#)

« 上一篇: [程序员必须知道的延迟时间](#)
» 下一篇: [广告引擎解析](#)

posted @ 2016-07-05 14:33 [23lalala](#) 阅读(1028) 评论(0) [编辑](#) [收藏](#)
[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C++/C#源码：大型实时仿真组态图形源码
- 【前端】SpreadJS表格控件，可嵌入系统开发的在线Excel
- 【推荐】码云企业版，高效的企业级软件协作开发管理平台
- 【推荐】程序员问答平台，解决您开发中遇到的技术难题

相关博文：

- [行存储和列存储](#)
- [转gridview的扩展，实现固定列和列头固定。](#)
- [SQL Server 2014聚集列存储索引](#)
- [列存储索引](#)
- [行存储与列存储笔记](#)