GREENPLUM DATABASE®

Brin Index在Greenplum 7中的理论与实现

—《Greenplum 7 新版本大剧透》系列直播

讲师：陈金豹，Greenplum内核研发

活动时间：2021年4月28日 20：00 - 21：00

GREENPLUM DATABASE®

Greenplum中文社区

https://cn.greenplum.org

博文 · 资料 · 文档 · 项目

GREENPLUM DATABASE®

微信技术讨论群
添加入群小助手：gp_assistant

微信公众号
技术干货、行业热点、活动预告

欢迎访问Greenplum中文社区：cn.greenplum.org

GREENPLUM DATABASE®

Brin Index 在Greenplum 7中的理论与实践

陈金豹，VMWare内核工程师

vmware®

Confidential | ©2021 VMware, Inc.

目录

Brin Index On Heap

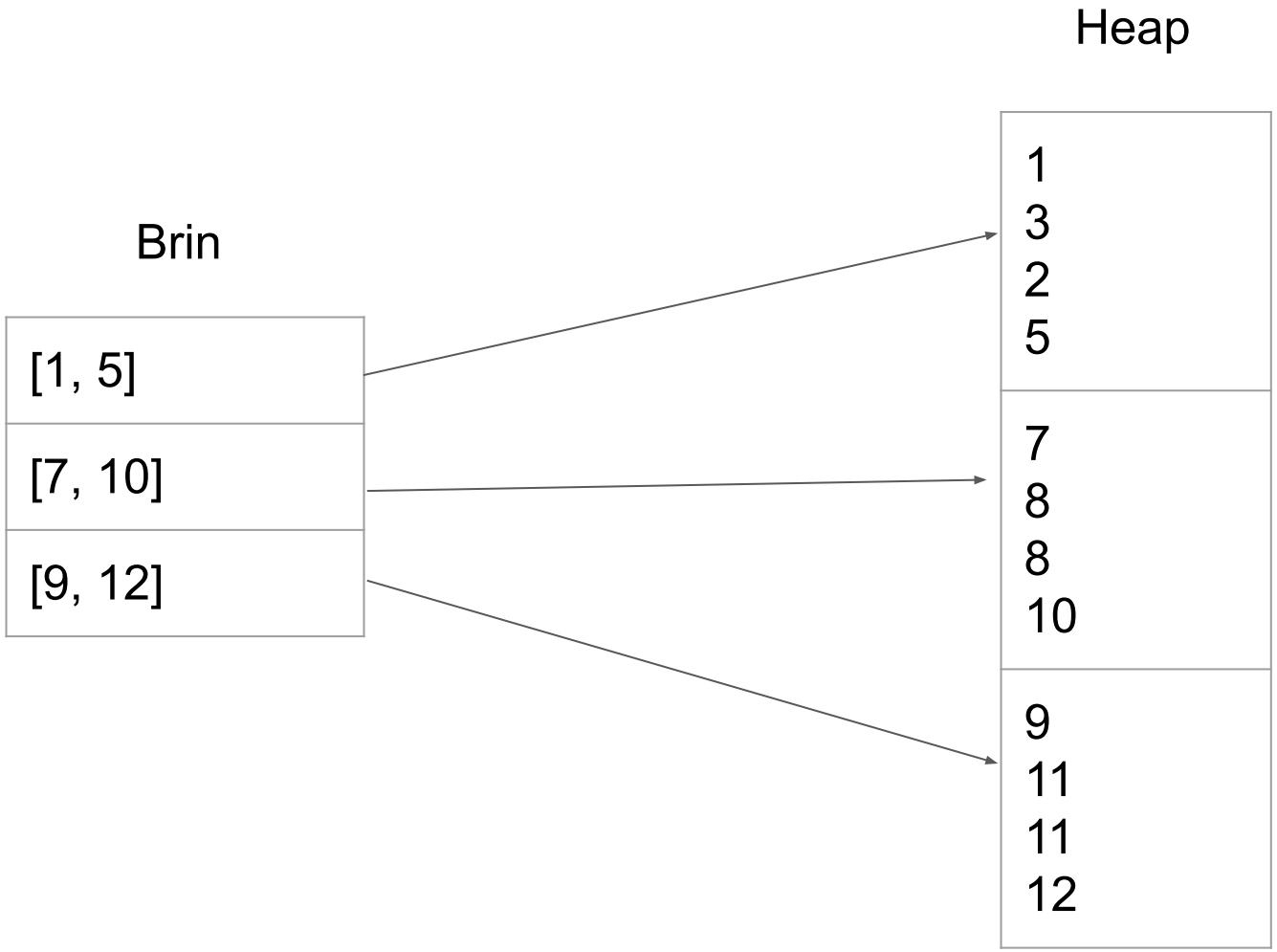
AppendOnly Table

Brin在AppendOnly Table上的实现

性能测试

Block Range Index

存储数据块中元组字段的最大最小值，用于过滤不符合条件的数据块



Confidential | ©2021 VMware, Inc.

Brin的优势和劣势

优势:

- 占用空间小
- 创建快

劣势:

- 只有在数据具有一定分布特点时才有用

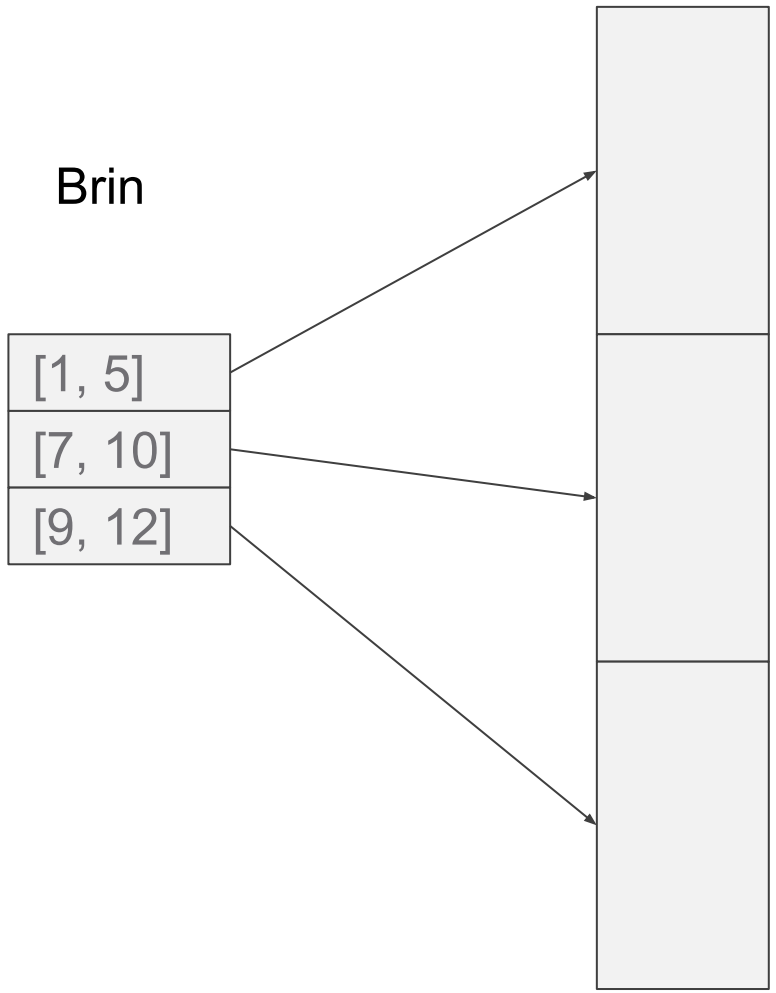
Brin的体积

Brin tuple: 20bytes

Block Range: 8K * 20 = 160K

Brin比Heap小8000倍

1 TB的Heap表只需要125M的Brin



Confidential | ©2021 VMware, Inc.

Brin的选择率

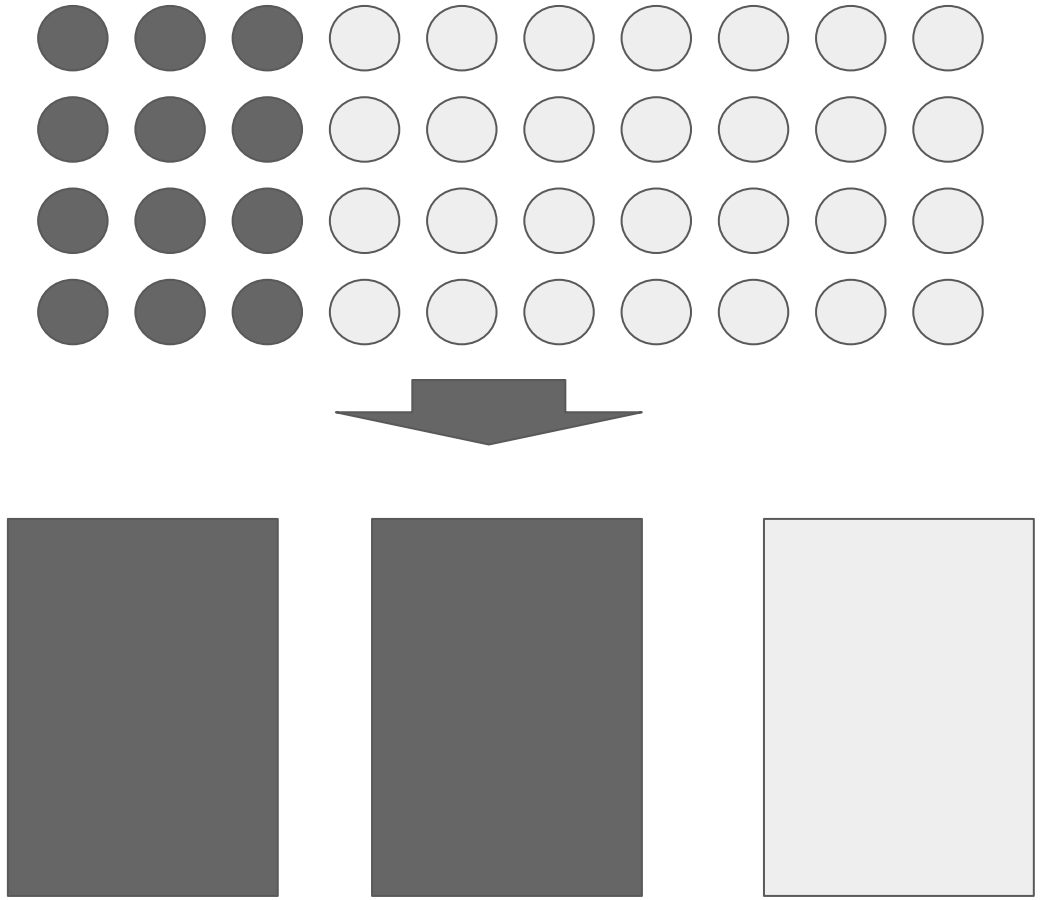
BlockNum: B = 1000

TupleNum: N = 1000000

TuplePerblock: M = 1000

Selection: a = 1%

Brin Selection: 1-((B-1)/B)^(N*a) = 1-0.000045



vmware Confidential | ©2021 VMware, Inc.

Brin的使用场景

GREENPLUM DATABASE®

表非常大

数据有一定的分布特性

我们不想再index上付出太多存储空间

Confidential | ©2021 VMware, Inc.

Brin Scan

select * from t where a > 1 and a < 8;

1, 3, 2, 5	7, 8, 8, 10	9, 11, 11, 12	10, 19, 11, 100

bit map

1 1 0 0

Insert Update Delele

每个元组代表一组Block每个字段的最大最小值

如果新的数据超出了最大最小值的范围，则更新元组

删除数据时不做任何操作

Confidential | ©2021 VMware, Inc.

Brin Vacuum

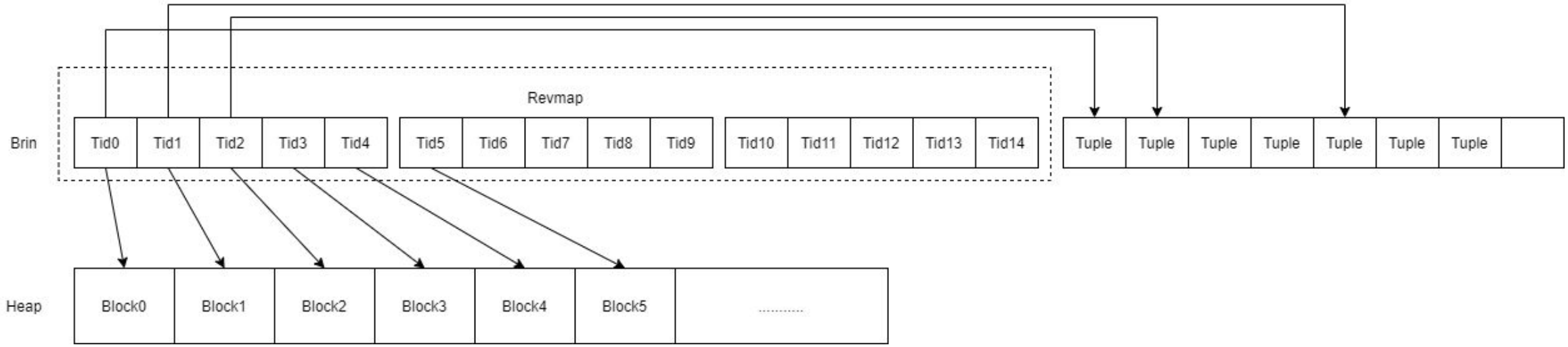
vacuum不做任何操作

vacuum full重建索引

Brin Storage

Revmap是一个逻辑上的数组，数组下标代表BlockNumber

Revmap的value是TID，指向存储有最大最小值的tuple



目录

Brin Index

AppendOnly Table

Brin在AppendOnly Table上的实现

性能测试

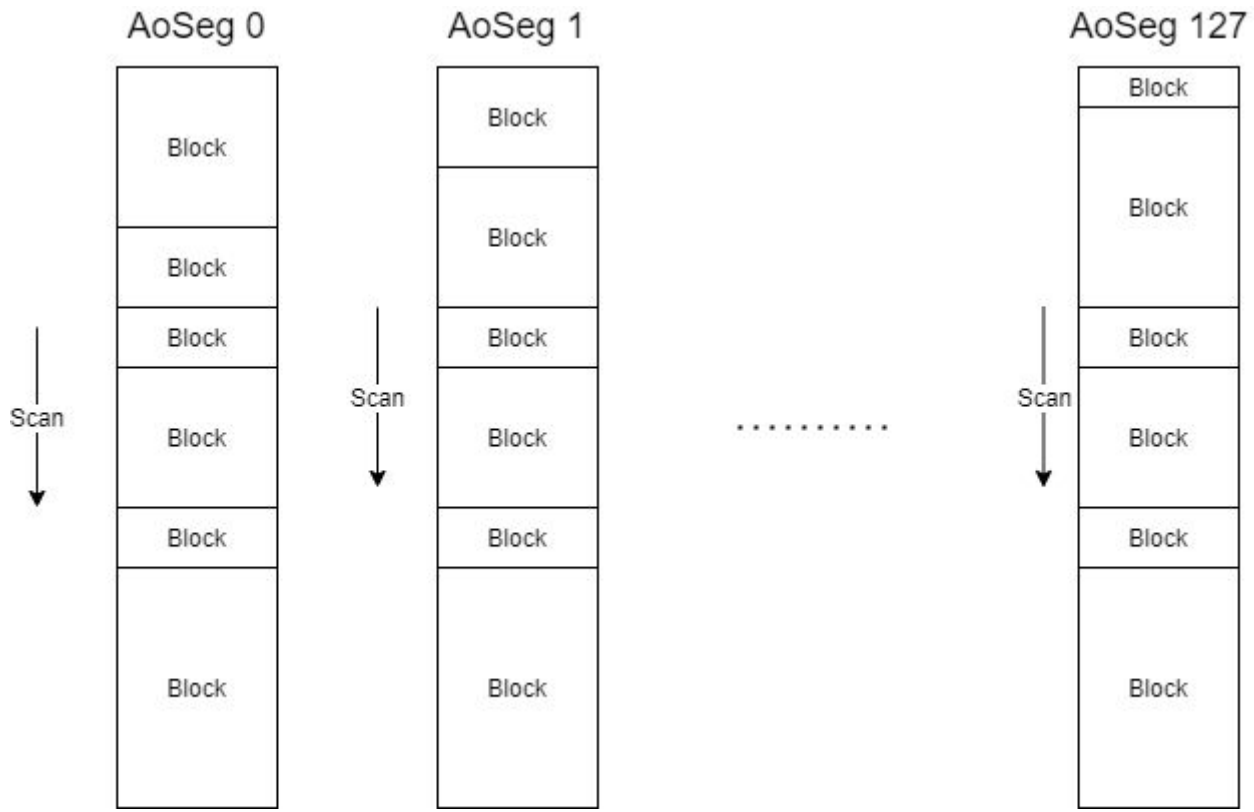
Confidential | ©2021 VMware, Inc.

AoTable

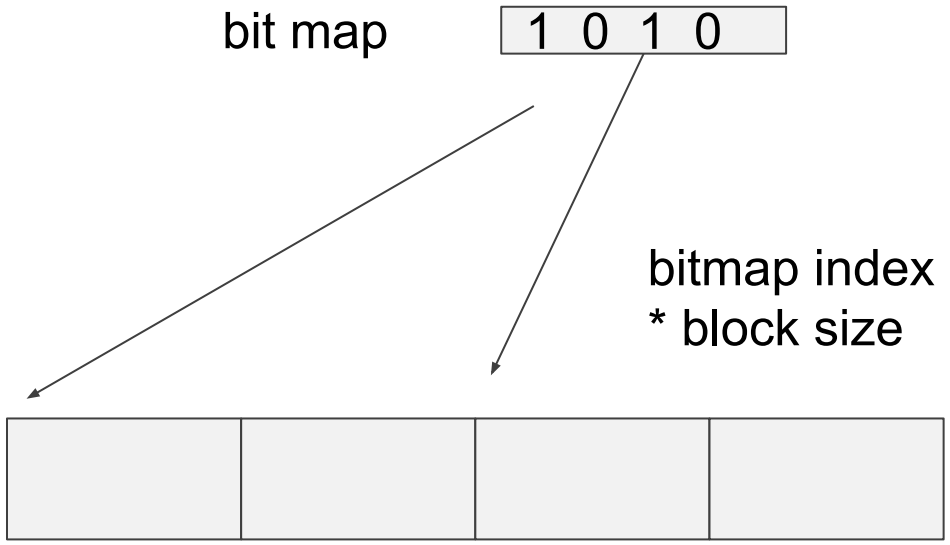
AppendOnly Table是一种紧凑的数据格式，适用于较少进行Update/Delete的场景

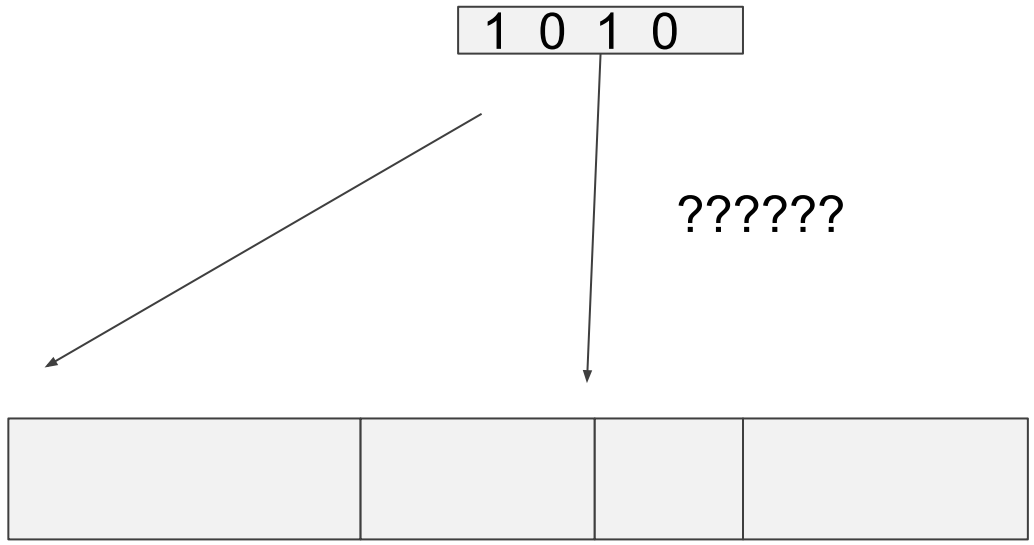
Tuple以紧凑的方式存储在变长的Block中，所以Block在写入磁盘后不能修改，只能向后追加新的Block

为了实现并发Insert，每个AO表逻辑上有128个AoSeg，每个事务向一个特定的AoSeg追加数据



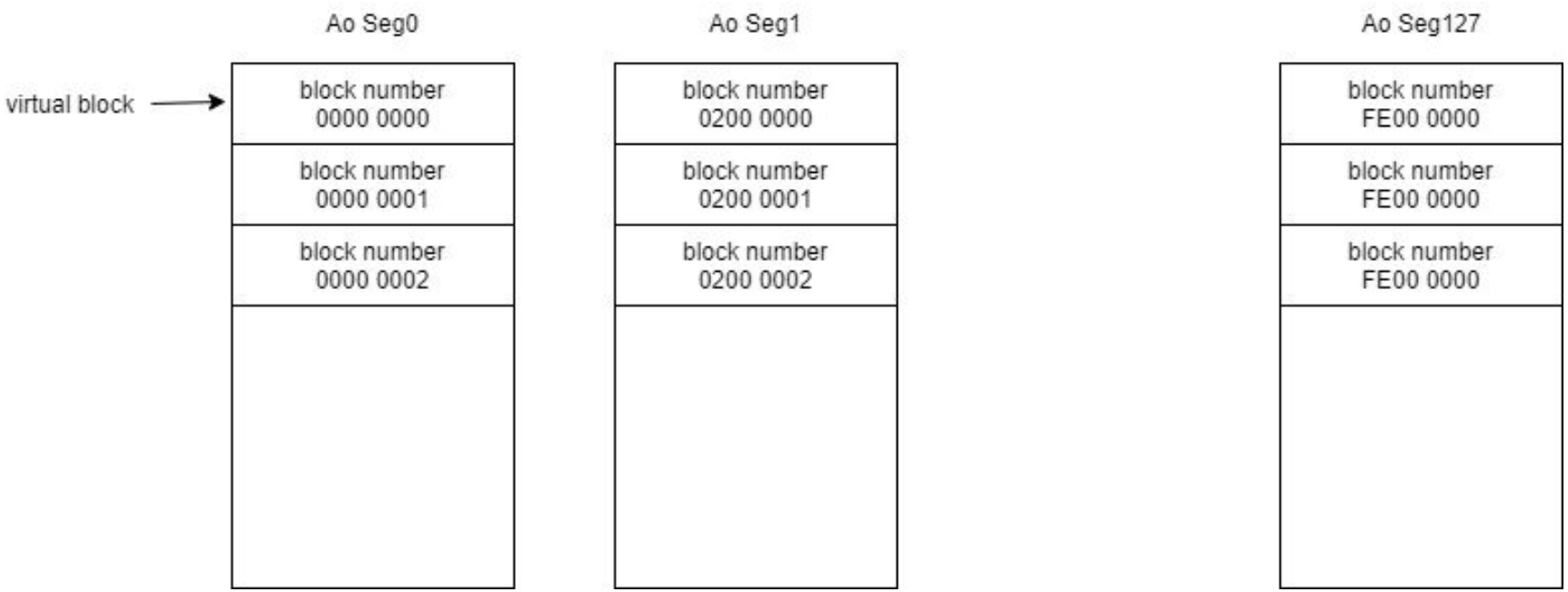
Brin Scan For Heap And AO





Confidential | ©2021 VMware, Inc.

Append only table



目录

Brin Index

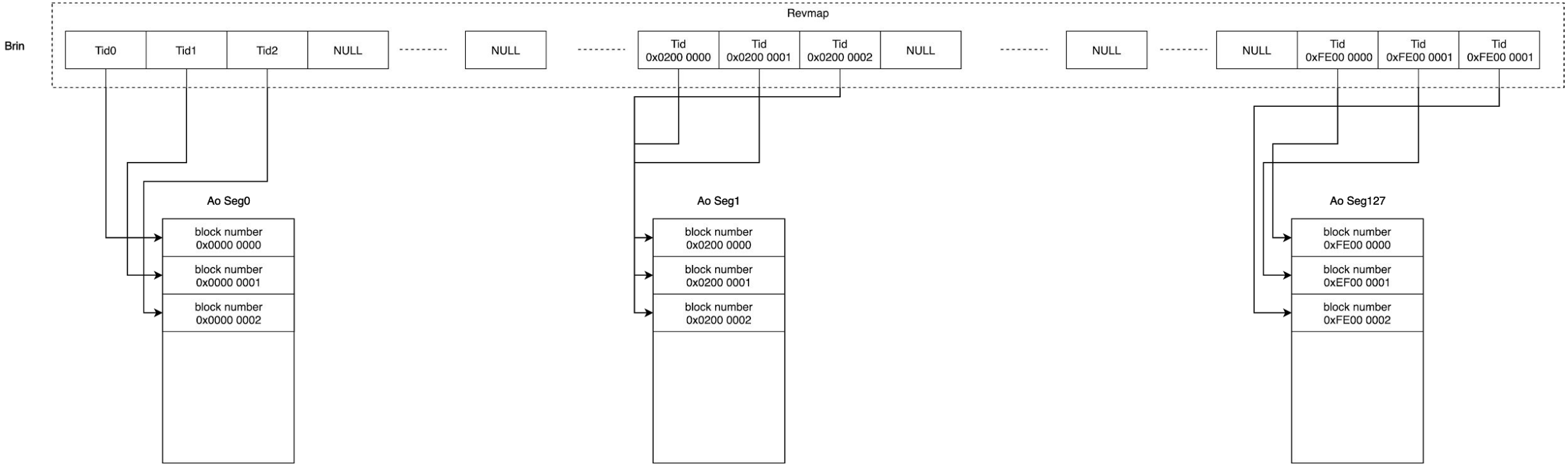
AppendOnly Table

Brin在AppendOnly Table上的实现

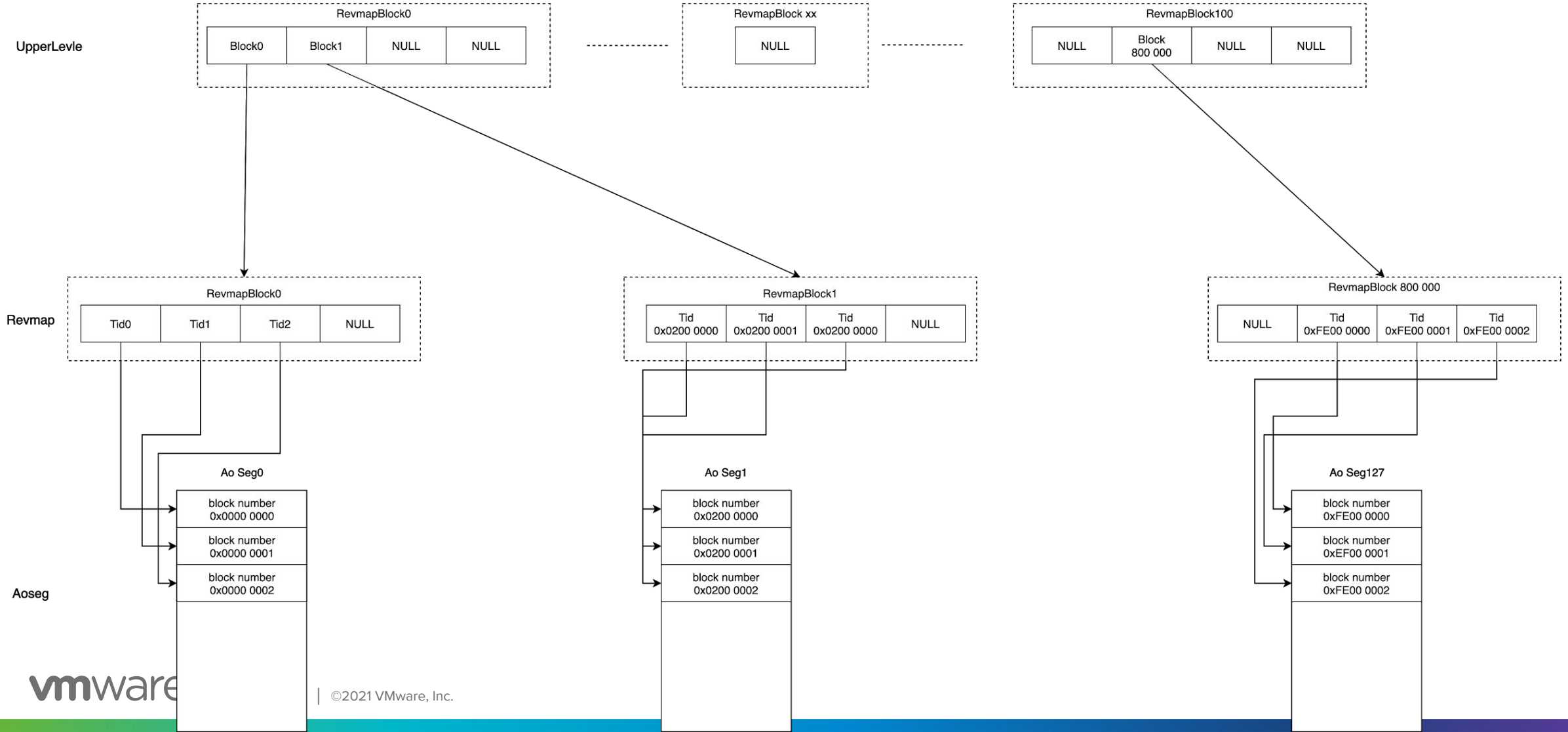
性能测试

Confidential | ©2021 VMware, Inc.

Revmap With AO



Add an upper level



©2021 VMware, Inc.

Brin on AO Storage

Upper Level：定长数据段，约为3.2MB

RevMap：变长数据段，大小随数据表数据量变化，向后增长

BrinTuples：存放Min Max的数据段，Page结构和Heap Page相同

UpperLevel (定长)	RevMap (变长)	BrinTuples (变长)

Confidential | ©2021 VMware, Inc.

Upper Level

number of records in the upper level

```
recordnumber = 2^32/TidNumPerPage
```

The corresponding relationship between the block number and the upper level array

```
upper_index=blocknum/TidNumPerPage
```

```
revmap_offset=blocknum%TidNumPerPage
```

目录

Brin Index

AppendOnly Table

Brin在AppendOnly Table上的实现

性能测试

Confidential | ©2021 VMware, Inc.

Performance Test

```
create table aocs(a int, b int) with(appendonly=true, ORIENTATION = column);

insert into aocs select i,i from generate_series(1,10000000)i;

create index abidx on ao using brin(b) with(pages_per_range=1);

create index atidx on ao using btree(b);
```

Test1

select gp_segment_id,* from ao where b=999999;

```
seqscan: 3957.205 ms

brin-bitmapscan: 36.456 ms

btree-bitmapscan: 18.111 ms
```

Confidential | ©2021 VMware, Inc.

Test2

select gp_segment_id,* from ao where b > 1000000 and b < 1010000;

```
seqscan: 5757.855 ms

brin-bitmapscan: 57.838 ms

btree-bitmapscan: 42.250 ms
```

Confidential | ©2021 VMware, Inc.

Test3

select gp_segment_id,* from ao where b > 1000000 and b < 2000000;

```
seqscan: 6413.329 ms

brin-bitmapscan: 2241.363 ms

btree-bitmapscan: 2141.896 ms
```

Confidential | ©2021 VMware, Inc.

空间占用

ao: 180,198,032

atidx-btree: 222,920,704

abidx-brin: 6,553,600

Confidential | ©2021 VMware, Inc.

GREENPLUM DATABASE®

Q & A

Confidential | ©2021 VMware, Inc.

GREENPLUM DATABASE®

微信技术讨论群
添加入群小助手：gp_assistant

微信公众号
技术干货、行业热点、活动预告

欢迎访问Greenplum中文社区：cn.greenplum.org

GREENPLUM DATABASE®

Thank You

Confidential | ©2021 VMware, Inc.