PCC POSTGRESCONF CN 2020 | PGConf.Asia 11.17-11.20

Greenplum的PG内核升级之路

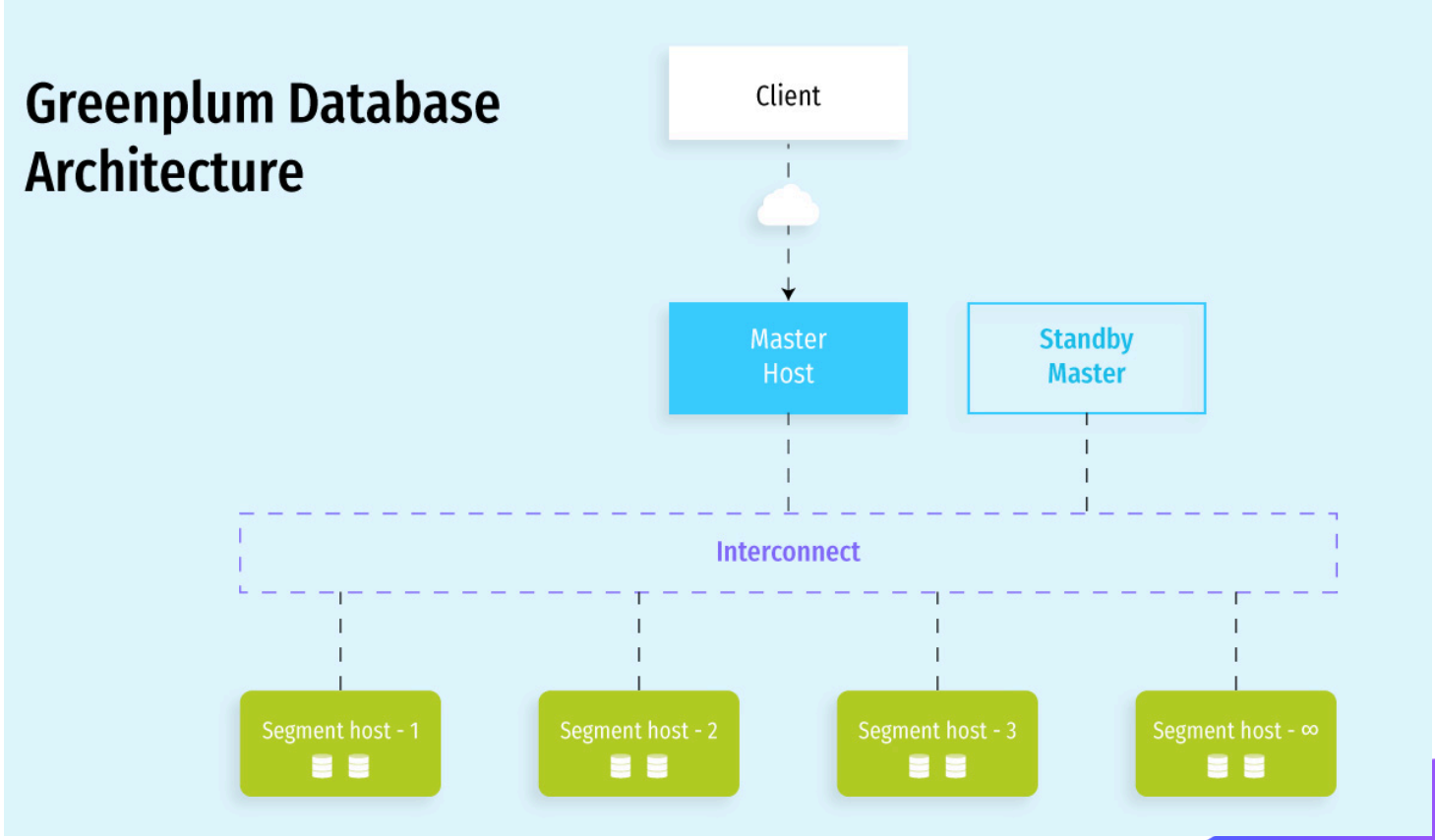
李晓亮 adlee@vmware.com

https://cn.greenplum.org
https://2020.postgresconf.cn

主要内容

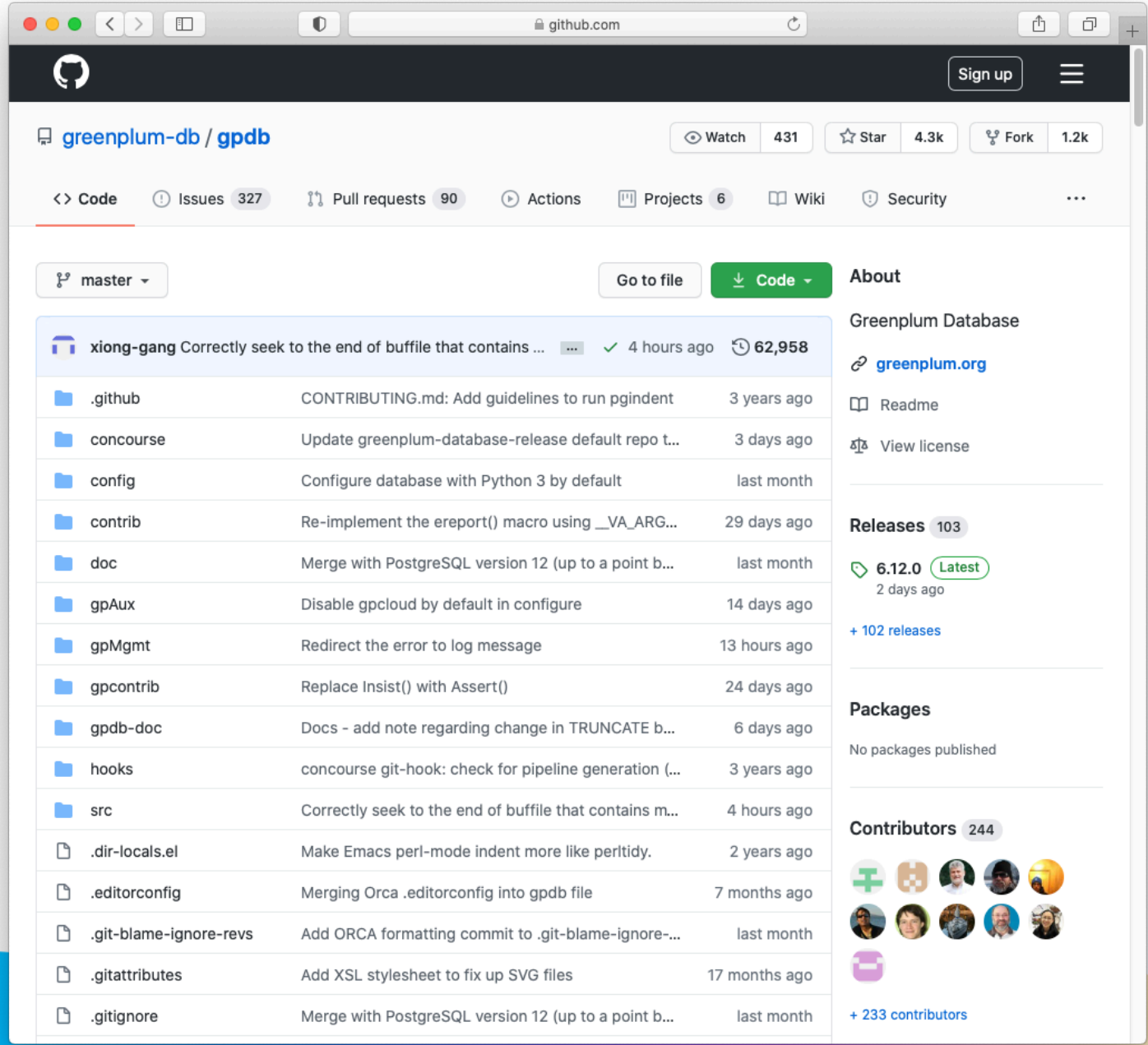
- Greenplum简介
- PG内核版本变化
- Greenplum如何升级PG内核
- PG内核升级过程中的挑战
- 目前PG内核升级迭代的进展

关于 Greemplum



关于 Greemplum

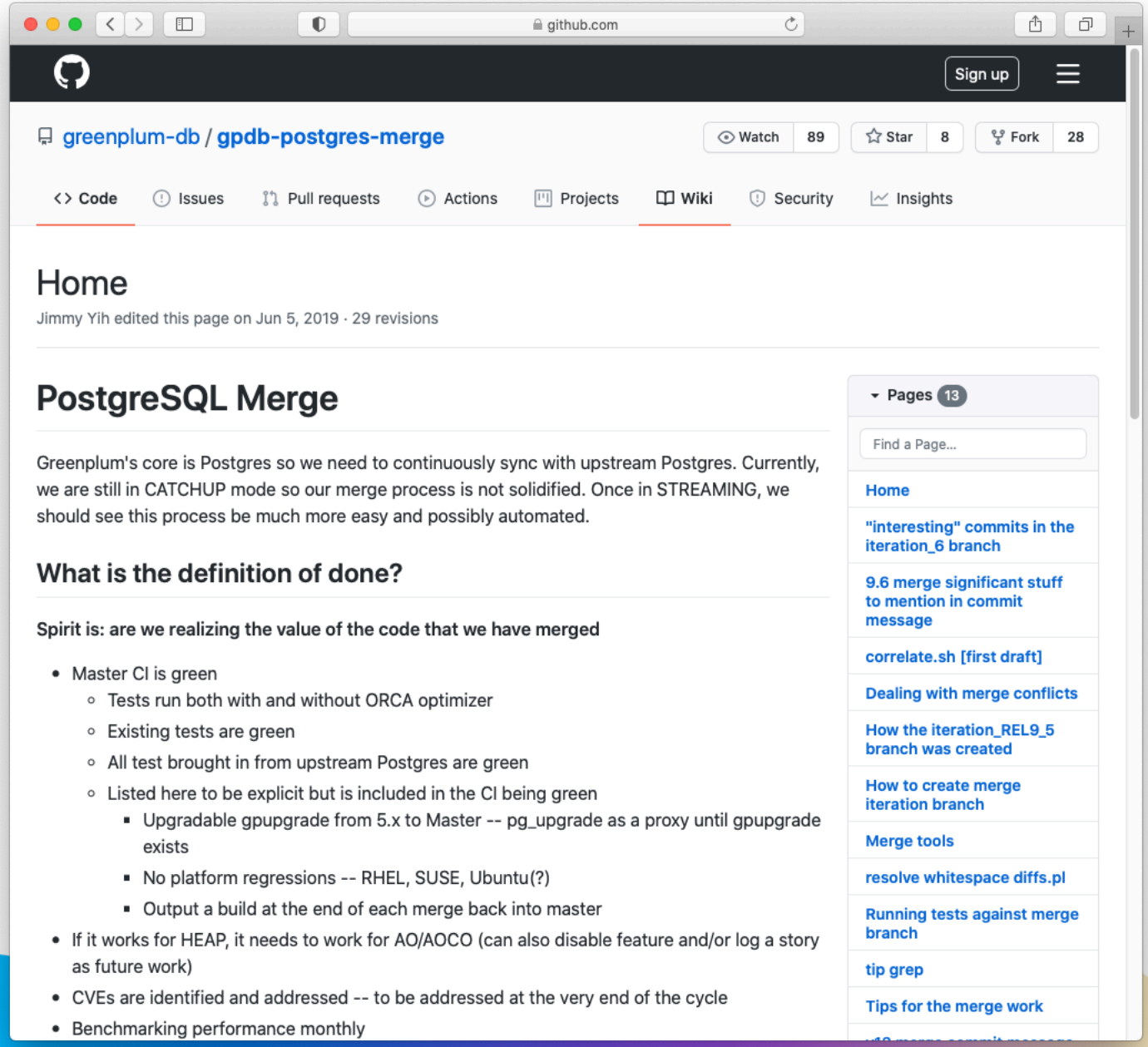
https://github.com/greenplum-db/gpdb



关于 Greemplum

- 4.3.x，闭源，基于PostgreSQL 8.2.15
- 5.x，开源，基于PostgreSQL 8.3.23
- 6.x，开源，基于PostgreSQL 9.4.24
- 7.0，开源，尚未发布，基于PostgreSQL 12

Greenplum如何升级PG内核



Greenplum如何升级PG内核

cherry-pick 还是 merge?

Greenplum如何升级PG内核

cherry-pick 还是 merge?

试过了，挨个cherry-pick的话一个迭代就能玩好几年。

还是merge一把梭吧。

Greenplum如何升级PG内核

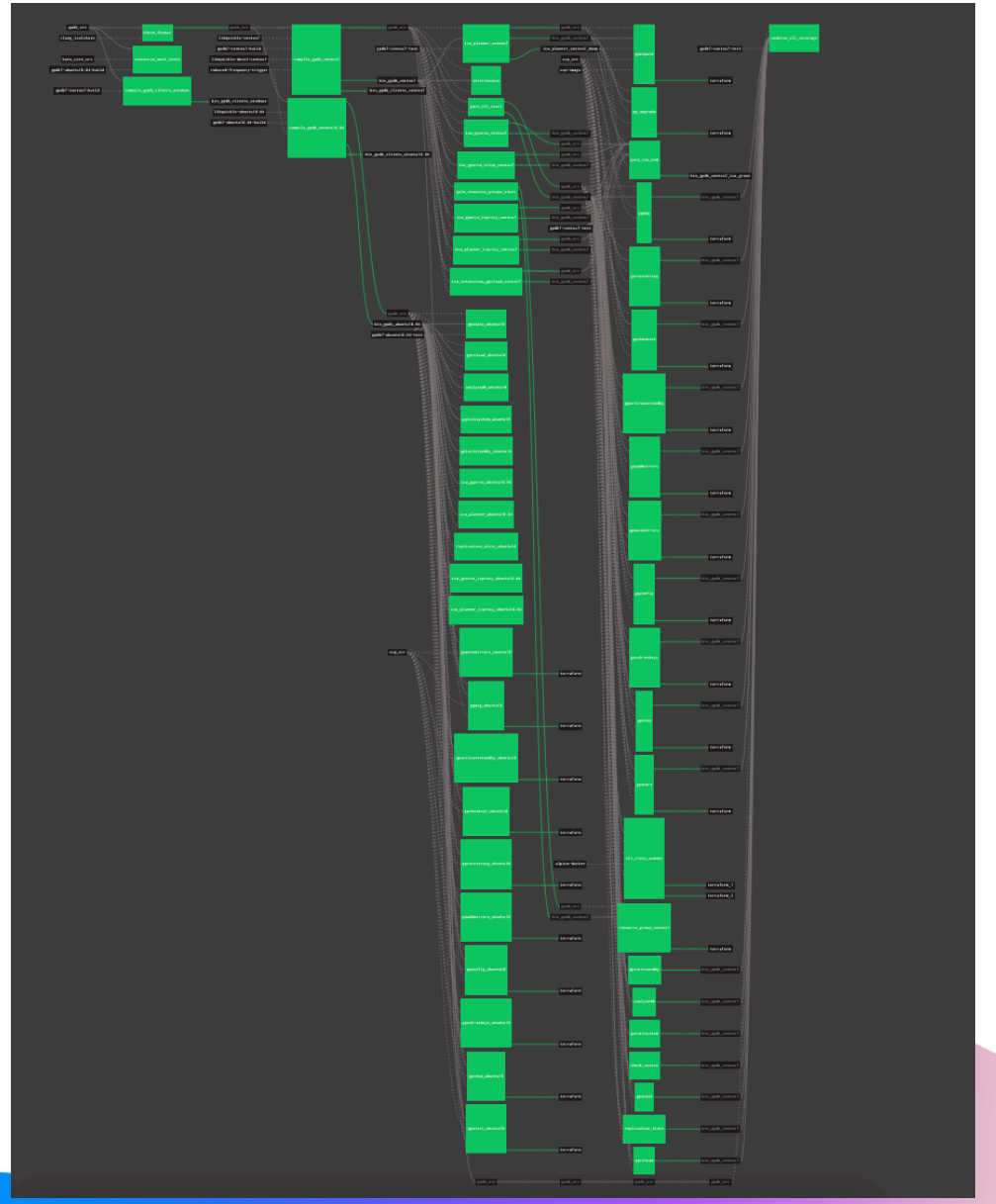
1，解决冲突使其能够编译

2，解决冲突使得initdb可以工作

3，解决冲突使得Greenplum可以启动运行

4，修正代码使得ICW和其它测试可以通过

Greenplum如何升级PG内核



PG内核升级过程中的挑战

1，冲突太多了

2，编译的时候互相依赖，很多时候是盲修

3，需要对内核代码很了解，需要对Greenplum的改动很了解

4，Greenplum的实现和上游PostgreSQL的实现之间的取舍

PG内核升级过程中的挑战

b5bce6c和9e1c9f9分别为PostgreSQL REL9_6_STABLE和 REL_12_STABLE与master branch的交汇点。

```
$ git lp b5bce6c..9e1c9f9|wc -l
    6527
$ git diff --stat b5bce6c..9e1c9f9
4521 files changed, 1116205 insertions(+), 891938 deletions(-)
```

PG内核升级过程中的挑战

9.6~12 代码冲突的数量：

$ grep -r “<<<<<<< HEAD” . | wc -l

5263

9.6~12 升级过程的工作量：

$ git lp 16c7a9db612..HEAD |wc -l

4436

PG内核升级过程中的挑战

```
diff --git a/src/backend/executor/nodeBitmapOr.c b/src/backend/executor/nodeBitmapOr.c
index 2bd7f00519..7f02a6bba3 100644
--- a/src/backend/executor/nodeBitmapOr.c
+++ b/src/backend/executor/nodeBitmapOr.c
@@ -149,64 +149,42 @@ MultiExecBitmapOr(BitmapOrState *node)
                Node       *subresult = NULL;

                /*
-                * We can special-case BitmapIndexScan children to avoid an explicit
-                * tbm_union step for each child: just pass down the current result
-                * bitmap and let the child OR directly into it.
+                * Note for further merge iteration:
+                *     Greenplum's BitmapIndexScan returns a StreamBitmap
                 */
-               if (IsA(subnode, BitmapIndexScanState))
-               {
-                       if (result == NULL) /* first subplan */
-                       {
-                               /* XXX should we use less than work_mem for this? */
-                               result = tbm_create(work_mem * 1024L,
-                                                                       ((BitmapOr *) node->ps.plan)->isshared ?
-                                                                       node->ps.state->es_query_dsa : NULL);
-                       }
+               subresult = MultiExecProcNode(subnode);

-                       ((BitmapIndexScanState *) subnode)->biss_result = result;
+               if (subresult == NULL)
+                       continue;

-                       subresult = (TIDBitmap *) MultiExecProcNode(subnode);
+               if (!(IsA(subresult, TIDBitmap) ||
+                         IsA(subresult, StreamBitmap)))
+                       elog(ERROR, "unrecognized result from subplan");

-                       if (subresult != result)
-                               elog(ERROR, "unrecognized result from subplan");
+               if (IsA(subresult, TIDBitmap))
+               {
+                       /* if it's a TIDBitmap, union into result */
+                       if (result == NULL)
+                               result = (TIDBitmap *)subresult;
+                       else
+                       {
+                               tbm_union(result, (TIDBitmap *)subresult);
+                               tbm_generic_free(subresult);
+                       }
                }
                else
                {
-                       /* standard implementation */
-                       subresult = (TIDBitmap *) MultiExecProcNode(subnode);
-
-                       if (subresult == NULL)
-                               continue;
-
-                       if (!(IsA(subresult, TIDBitmap) ||
-                                 IsA(subresult, StreamBitmap)))
-                               elog(ERROR, "unrecognized result from subplan");
-
-                       if (IsA(subresult, TIDBitmap))
+                       /* if it's a StreamBitmap, union into node->bitmap */
+                       if (node->bitmap)
                        {
-                               /* if it's a TIDBitmap, union into result */
-                               if (result == NULL)
-                                       result = (TIDBitmap *)subresult;
-                               else
-                                       tbm_union(result, (TIDBitmap *)subresult);
+                               if (node->bitmap != subresult)
+                                       tbm_union(result, (TIDBitmap *)subresult);
```

PG内核升级过程中的挑战

```
Author: Adam Lee <ali@pivotal.io>
Date:   Tue May 26 15:57:53 2020 +0800

    Move to next block if rs_cindex == INT16_MAX + 1 and it's a lossy page

    TID offsets are not zero based, INT16_MAX + 1 is a valid offset number.

    Don't return false if rs_cindex == INT16_MAX and pseudoHeapOffset is
    INT16_MAX + 1, but at the next one.

diff --git a/src/backend/access/aocs/aocsam_handler.c b/src/backend/access/aocs/aocsam_handler.c
index 5e0129ed0e..9e9f3e652a 100644
--- a/src/backend/access/aocs/aocsam_handler.c
+++ b/src/backend/access/aocs/aocsam_handler.c
@@ -1376,7 +1376,7 @@ aoco_scan_bitmap_next_tuple(TableScanDesc scan,
                 */
                if (tbmres->ntuples == -1)
                {
-                       if (aocoscan->rs_cindex == INT16_MAX)
+                       if (aocoscan->rs_cindex == INT16_MAX + 1)
                                return false;

                        /*
@@ -1422,6 +1422,7 @@ aoco_scan_bitmap_next_tuple(TableScanDesc scan,
                                ExecClearTuple(slot);
                }
                else
+               {
                        if(aocs_fetch(aocoscan->aocofetch, &aoTid, slot))
                                ExecStoreVirtualTuple(slot);
                        else
@@ -1429,6 +1430,7 @@ aoco_scan_bitmap_next_tuple(TableScanDesc scan,
                                if (slot)
                                        ExecClearTuple(slot);
                        }
+               }

                if (TupIsNull(slot))
                        continue;
```

PG内核升级过程中的挑战

Support GPDB partition syntax for PG12 merge

TASKS (69/69)

目前PG内核升级迭代的进展

Merge with PostgreSQL v12 #10862

Merged

hlinnaka merged 6,529 commits into `greenplum-db:master` from `hlinnaka:iteration_REL_12`

Sep 22, 2020

Conversation 11 Commits 250 Files

5,133 changed files

822,617 additions and 746,216 deletions

目前PG内核升级迭代的进展

```
postgres=# select version();
                                                                                                version
-------------------------------------------------------------------------------------------------------------------------------------------------
 PostgreSQL 12beta2 (Greenplum Database 7.0.0-alpha.0+dev.14027.g4dc25ad70c build dev) on x86_64-pc-linux-gnu, compiled by gcc (GCC) 4.8.5 20150623 (Red Hat 4.8.5-39), 64-bit compiled on Nov  2 2020 14:15:05 (with assert checking)
(1 row)
```

PG 12带来的新特性

1. 基于上游的分区表实现

之前Greenplum的分区表实现代码已经全部被替换为上游PostgreSQL的实现，同时实现了一个中间层以同时支持原来的和上游的分区表语法。

2. Access method API

PostgreSQL借这套API极大地提升了存储扩展能力，实现一个新的存储方法会变得更规范、方便和清晰。Greenplum在此次升级中也和上游一致，全部转换为此API，包括特有的Append-only行存和列存。

PG 12带来的新特性

性能和稳定性的提升
Generated columns语句的支持
Index-only scan的增强
哈希索引的日志支持
SQL/JSON的增强
GSSAPI验证
LDAP服务支持
多重要素验证的支持
CREATE PROCEDURE和CALL语句
COPY FROM … WHERE
FDW下推的增强

PG 12带来的新特性

有关 PostgreSQL 内核的更多更新，请参考官方文档。

https://www.postgresql.org/docs/release/10.0/

https://www.postgresql.org/docs/release/11.0/

https://www.postgresql.org/docs/release/12.0/

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®

CONTACT

THANKS

李晓亮 adlee@vmware.com

CHINA POSTGRESQL ASSOCIATION | PCC POSTGRESCONF CN 2020 | PGConf.Asia