SQL聚类扩展语句Cluster-by：PostgreSQL实现

孙鹏
bluevaley@gmail.com
中科院软件所

2011-7-15

提纲

- 本文目的
- 研究动机
- SQL聚类扩展语句Cluster-by语法语义
- 如何在PostgreSQL上实现？
- 其它话题
- 致谢

提纲

- 本文目的
- 研究动机
- SQL聚类扩展语句Cluster-by语法语义
- 如何在PostgreSQL上实现？
- 其它话题
- 致谢

本文目的

- 在科研中，使用PostgreSQL来实验验证一个新想法或新方法是很常见的，本文借助于“Cluster-by语句PostgreSQL实现”这样一个主题，愿与高年级本科生、研究生以及对PostgreSQL内核感兴趣的技术人员一起分享如何建立环境去调试PostgreSQL、如何去理解其内核、如何增加自己的代码等等，让您不再畏惧PostgreSQL内部机制，建立信心，并能够独立的开展工作。

提纲

- 本文目的
- 研究动机
- SQL聚类扩展语句Cluster-by语法语义
- 如何在PostgreSQL上实现？
- 其它话题
- 致谢

研究动机

- 研究动机：在对聚类挖掘研究中我们发现：
  1. 当前SQL (Structured Query Language)语言标准中缺少统一的支持空间或非空间聚类挖掘的语法语义
  2. 在数据库中（in-database）进行聚类有许多好处，例如不需要在数据库和外部聚类模块之间传送大量的数据，减少了网络负载
  3. 从对数据进行分组的角度来看，聚类是一类模糊的分组，是对Group-by语句的补充

提纲

- 本文目的
- 研究动机
- SQL聚类扩展语句Cluster-by语法语义
- 如何在PostgreSQL上实现？
- 其它话题
- 致谢

SQL语言聚类扩展：Cluster-by语句

- Cluster-by语句的SQL语法定义：

  <table expression> ::= <from clause>

  [<where clause>]

  [<group by clause>|**<Cluster by clause>**]

  [<having clause>]

  [<window clause>]

  <Cluster by clause> ::= CLUSTER BY <clustering column reference list>

  [USING <clustering algorithm function>]

SQL语言聚类扩展：Cluster-by语句

- Cluster-by语句的SQL语义：主要针对多个字段进行聚类时进行了语义解释，若有：

  CLUSTER BY $w_1*col_1,w_2*col_2,...,w_m*col_m$

  USING CAF(p_1, p_2,...,p_n)

  则加和距离函数ODF，定义为：

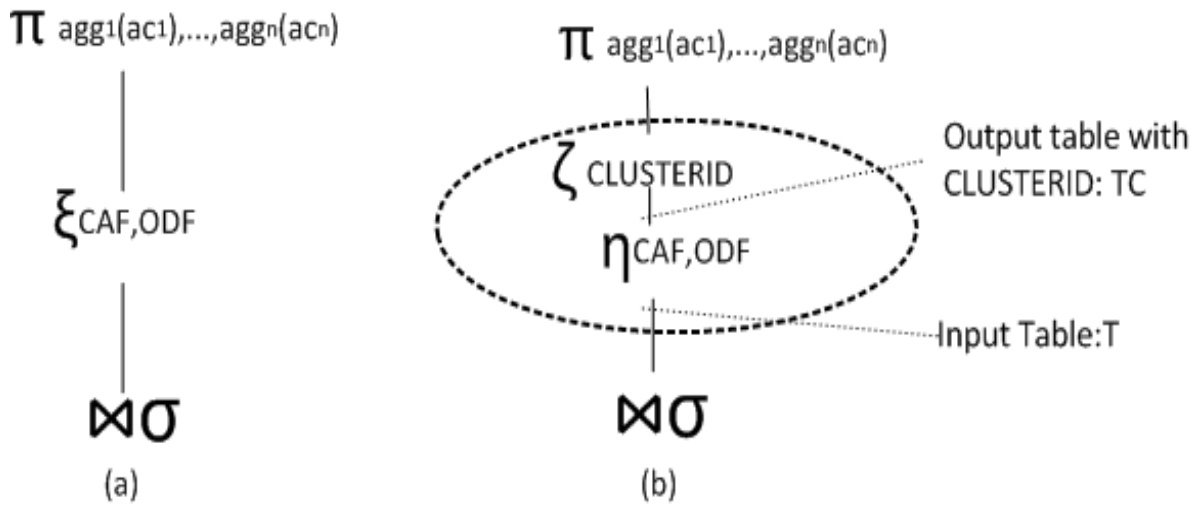
  $ODF = \Sigma (w_i*df_i(col_i)) \mid i = 1,...,m$

  则在ODF上执行聚类函数CAF获得g组簇定义为：

  $\xi_{A,CAF,ODF} = \{ M_i \mid M_i \subseteq A, M_i$由CAF使用ODF生成$,i=1,...,g \}$

SQL语言聚类扩展：Cluster-by语句

- Cluster-by的Plan tree节点：



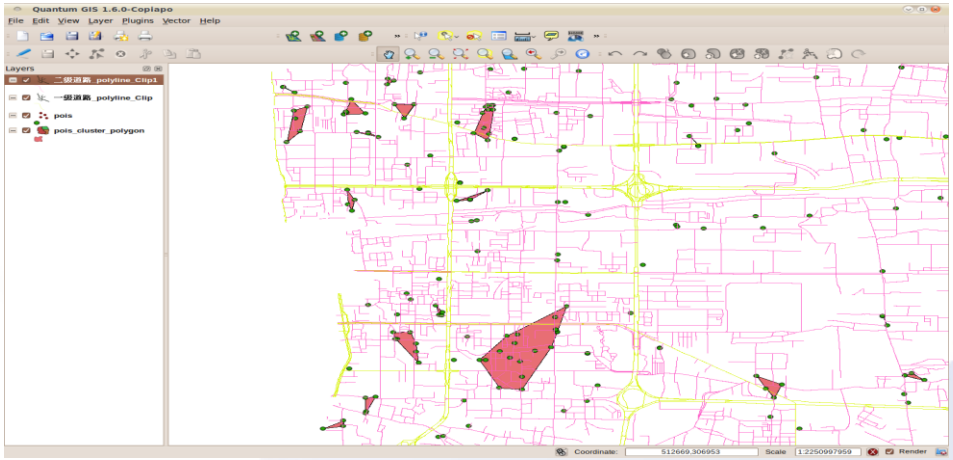
SQL语言聚类扩展：一个使用示例

- Cluster-by的一个使用示例：

```
postgres@postgres-laptop: ~
File  Edit  View  Terminal  Help
postgres@postgres-laptop:~$ psql mydb;
psql (8.4.2)
Type "help" for help.

mydb=# INSERT INTO pois_cluster_polygon
mydb-# SELECT St_geomfromtext(St_astext(St_convexhull(St_collect(location))))
mydb-# FROM   chaoyang
mydb-# CLUSTERING BY location USING DBSCAN(3,300.0);
INSERT 0 64
mydb=#
```

在chaoyang表上执行*Cluster-by*语句



使用Quantum GIS查看生成的凸包

提纲

- 本文目的
- 研究动机
- SQL聚类扩展语句Cluster-by语法语义
- 如何在PostgreSQL上实现？
- 其它话题
- 致谢

PostgreSQL内实现：搭建调试环境

- 下面列出了两篇关于环境搭建的文章，分别针对Red Hat和Ubuntu两类操作系统：
  1. rhel6+postgresql8.4+postgis1.4+eclipse CDT3.6 调试环境搭建
  2. postgresql 8.4 +postgis1.5 +eclipse CDT3.6 + ubuntu 9.10 +vmware 6.0 开发调试环境搭建

  主要参考：http://wiki.postgresql.org/wiki/Working_with_Eclipse

PostgreSQL内实现：基础

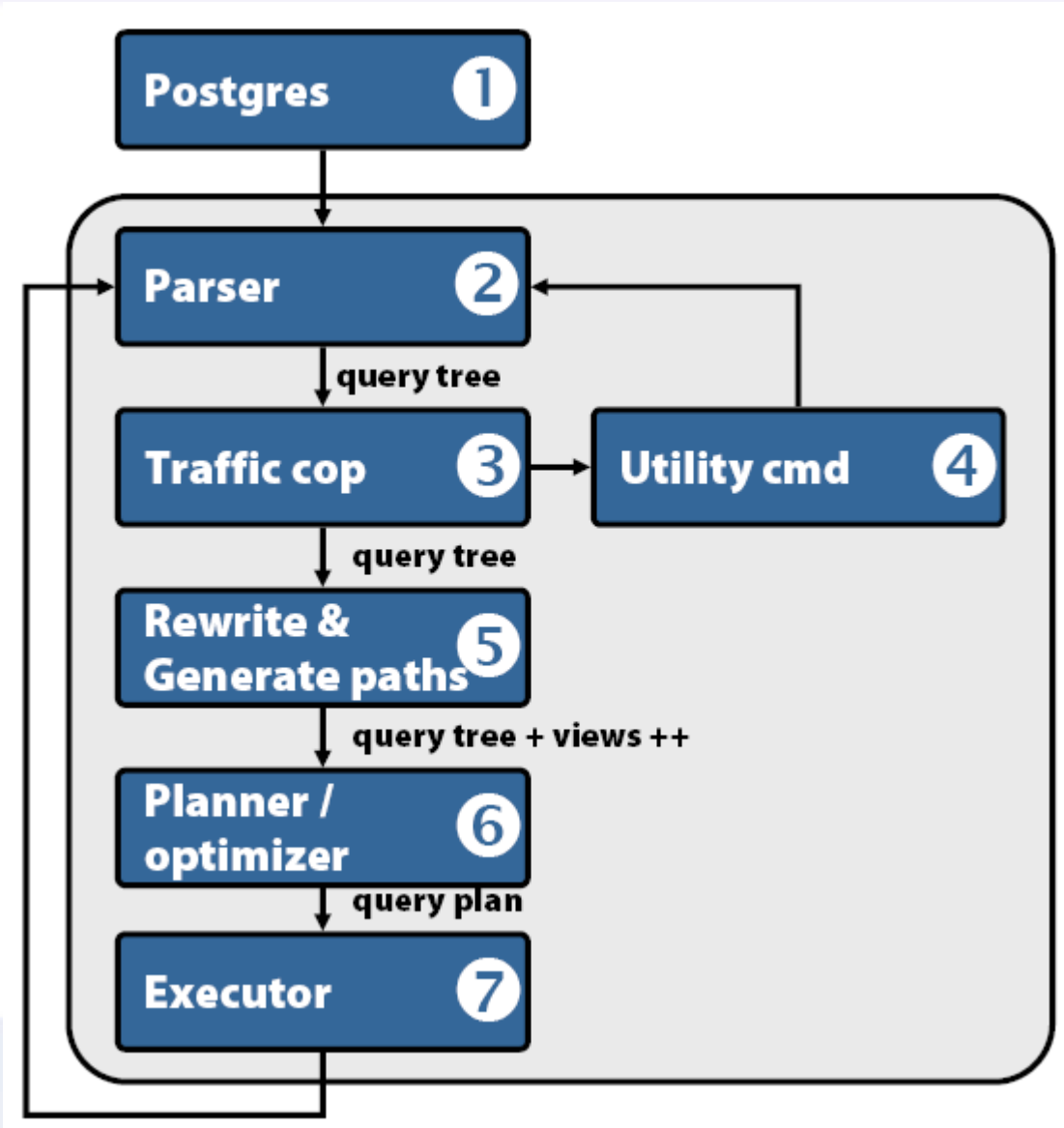
- 下面列出了一些参考资料：
  1. PostgreSQL Internals Through Pictures Bruce’s blog
  2. PostgreSQL 9.1,9.0, 8.4… Documents
  3. Postmaster的Shared Memory中的shmem index table 内存结构
  4. Postmaster的Shared Memory中的shared buffer pool内存结构
  5. Postmaster的Memory Context 初始化内存结构
  6. PostgresMain()中重要的几个初始化
  7. postgresql中parse tree内存结构
  8. postgresql中plantree内存结构
  9. 一天之内不再畏惧lex&yacc之必备参考资料

  ...

PostgreSQL内实现

- 我们先看看需要修改哪些阶段：

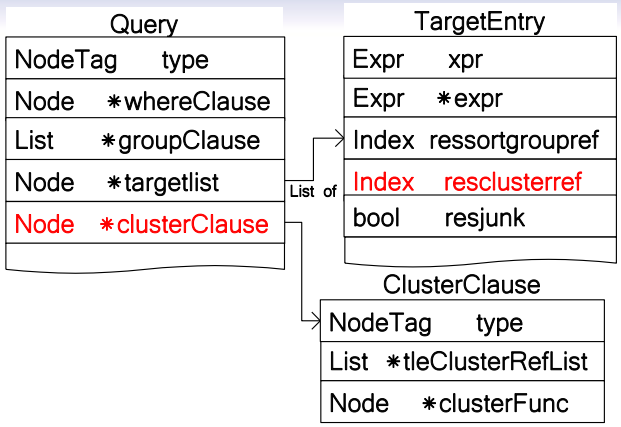
  Get To Know PostgreSQL，摘自Bruce's 主页



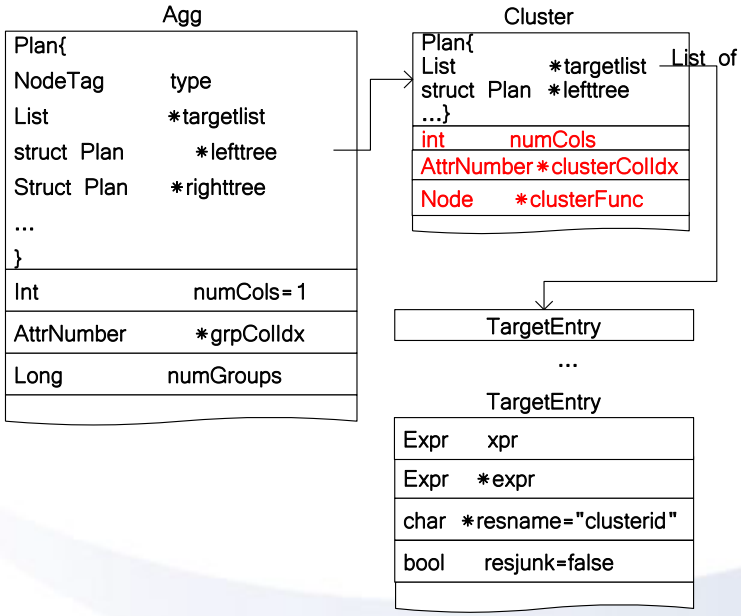
PostgreSQL内实现：主要数据结构

```
cluster_clause:
CLUSTERING BY expr_list cluster_func_el{
  $$ = makeNode(ClusterVar);
  $$->columnExpList = $3;
  $$->funcCall = $4;
  $$->location = @1;
 } | /*EMPTY*/ { $$ = NIL; } ;
cluster_func_el:
 USING cluster_func_expr { $$ = $2; }
 | /*EMPTY*/            { $$ = NIL; };
Cluster_func_expr: func_expr  { $$ = $1;};
```

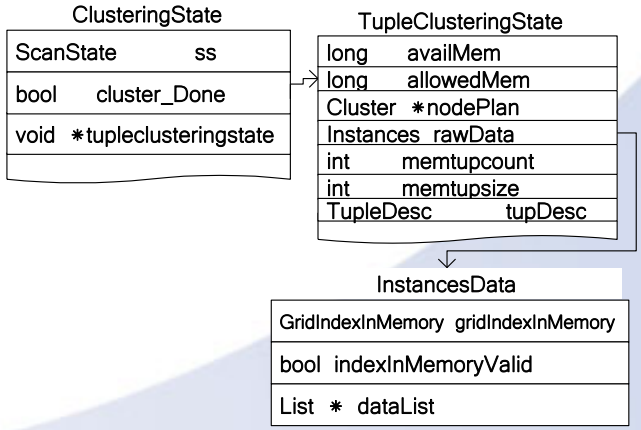
（a） cluster_clause definition in gram.y



（b） Query



（c） Optimizer

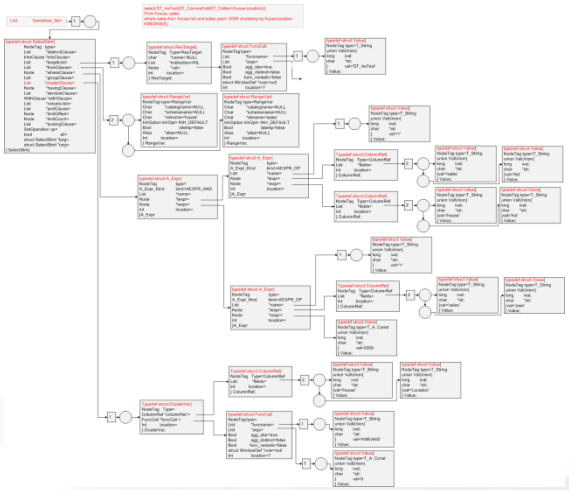


（d） Executor

PostgreSQL内实现：一个示例

- 我们以如下语句示例，看看到底Parse Tree，Query Tree，Plan Tree等等是什么样子？

```
select ST_AsText(ST_ConvexHull(ST_Collect (house.location)))
From house, sales
where sales.hid= house.hid and sales.year= 2009 clustering by
house.location KMEANS(5);
```



- 点击VISIO大图

PostgreSQL内实现：聚类算法的执行

- 我们以部分代码简单说明一下

```
TupleTableSlot *ExecCluster(ClusteringState *node) {
    …
    if (!node->cluster_Done) {
        TupleTableSlot *slot;
        for (;;) {
            slot = ExecProcNode(outerNode); ….
            tupleCluster_puttupleslot(tupleclusteringstate, slot);…
        }
       clusteringDispatchInMemory(tupleclusteringstate); //聚类
    } …
    bool gotit = false;
    gotit = tupleCluster_gettupleslot(tupleclusteringstate, slot);
    if (gotit) {return slot;}else{return NULL}
}
```

提纲

- 本文目的
- 研究动机
- SQL聚类扩展语句Cluster-by语法语义
- 如何在PostgreSQL上实现？
- 其它话题

- 致谢

其它话题

- 在研究过程中，我们发现：
  1. 在数据库多查询优化中发现，前面的聚类挖掘部分结果可以被后续的挖掘请求所使用：多查询优化
  2. 在采用PostgreSQL实现聚类扩展中发现，可以利用多进程的机制，充分挖掘当前主流的多核平台提供的强有力的并行计算能力：多核并行算法

Q&A

谢谢大家

感谢北京才冠软件有限公司技术人员的绘图