开源大数据引擎：Greenplum 数据库架构分析

姚延栋
yyao@pivotal.io
Pivotal研发总监

Pivotal

日程

- Greenplum 数据库简介
- Greenplum 数据库架构
- Greenplum 数据库核心组件
- Greenplum 数据库SQL执行过程
- Greenplum 数据库开源

Greenplum 简介

Pivotal

大数据 ≠ Hadoop

大多情况下，GPDB 更适合做大数据存储、计算、分析和挖掘的引擎。

GPDB：为大数据存储、计算、挖掘而设计

- 标准 SQL 数据库：ANSI SQL 2008 标准，OLAP，JDBC/ODBC
- 支持ACID、分布式事务
- 分布式数据库：线性扩展，支持上百物理节点
- 企业级数据库：全球大客户超过 1000+ 安装集群
- 百万行源代码，超过10年的全球研发投入
- 开源数据库(greenplum.org)，良性生态系统

客户


Ford
Go Further


GE


SAMSUNG


中信银行
CHINA CITIC BANK


ADP
Cobalt


AMERICAN FAMILY
INSURANCE


BOEING


Walmart


NYSE Euronext


Regence


INTTRA


FICO


symantec.


Deutsche Bank


IRS


McAfee


CareCore
NATIONAL


FRANKLIN TEMPLETON
INVESTMENTS


Bank 中国光大银行
CHINA EVERBRIGHT BANK


SUNGARD


comScore


TRAVELERS


T··Mobile·


ORBITZ


HAVAS
DIGITAL


skype


SONY


NASDAQ


USASPENDING.GOV

合作伙伴


sas


vmware


MicroStrategy
Best In Business Intelligence


COGNOS


JASPER SOFT


INFORMATICA
The Data Integration Company


SnapLogic


talend
open data solutions


ALPINE MINER


Datameer
Powerfully Simple


SAP


pentaho


+ableau
SOFTWARE


CSC


KARMASPHERE

Greenplum 架构

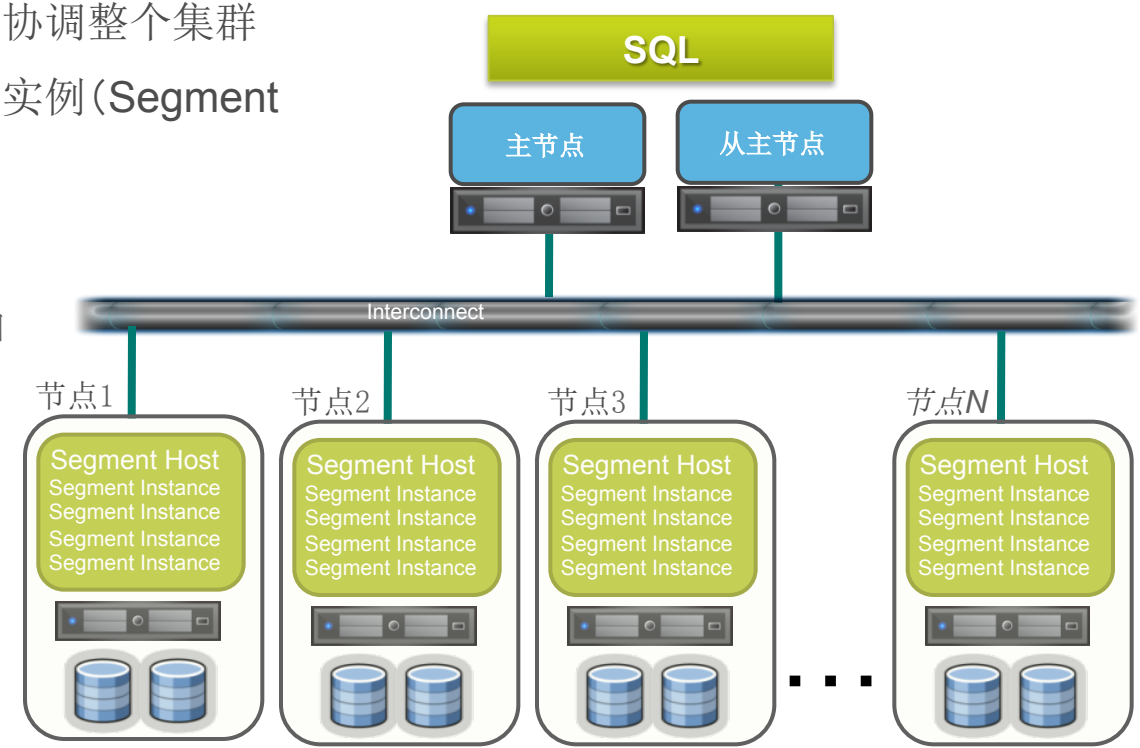
Pivotal

平台概况


客户端访问和工具
客户端访问
ODBC, JDBC, OLEDB, etc.
第三方工具
BI 工具，ETL 工具
文本分析，数据挖掘等
管理工具
GP Command Center
GP Workload Manager
产品特性
加载 & 数据联邦
高速数据加载
近实时数据加载
任意系统数据访问
存储 & 数据访问
混合存储引擎（行存&列存）
多种压缩，多级分区表
索引（B树，位图，GiST）
安全性
语言支持
标准SQL支持，SQL 2003 OLAP扩展
支持 MapReduce
扩展编程语言
(Python,R, Java, Perl, C)
服务
多级容错机制
在线系统扩展
任务管理
核心MPP
架构
无共享大规模并行处理
先进的查询优化器
多态存储系统
并行数据流引擎
高速软数据交换机制
MPP Scatter/Gather 流处理

MPP 无共享体系

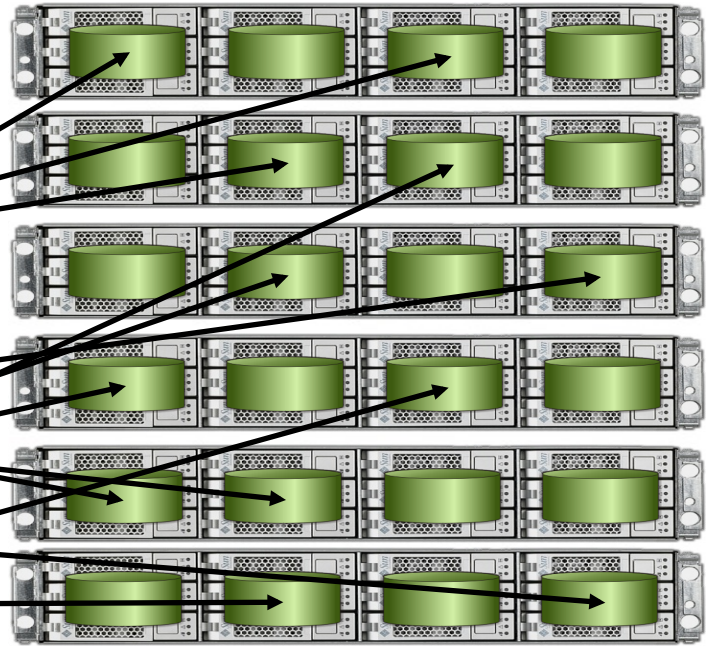
- 主节点和从主节点，主节点负责协调整个集群
- 一个数据节点可以配置多个节点实例(Segment Instances)
- 节点实例并行处理查询（SQL）
- 数据节点有自己的CPU、磁盘和内存(Share nothing)
- 高速Interconnect处理持续数据流(Pipelining)



数据分布：并行化的根基

最重要的策略和目标是均匀分布数据到各个数据节点。

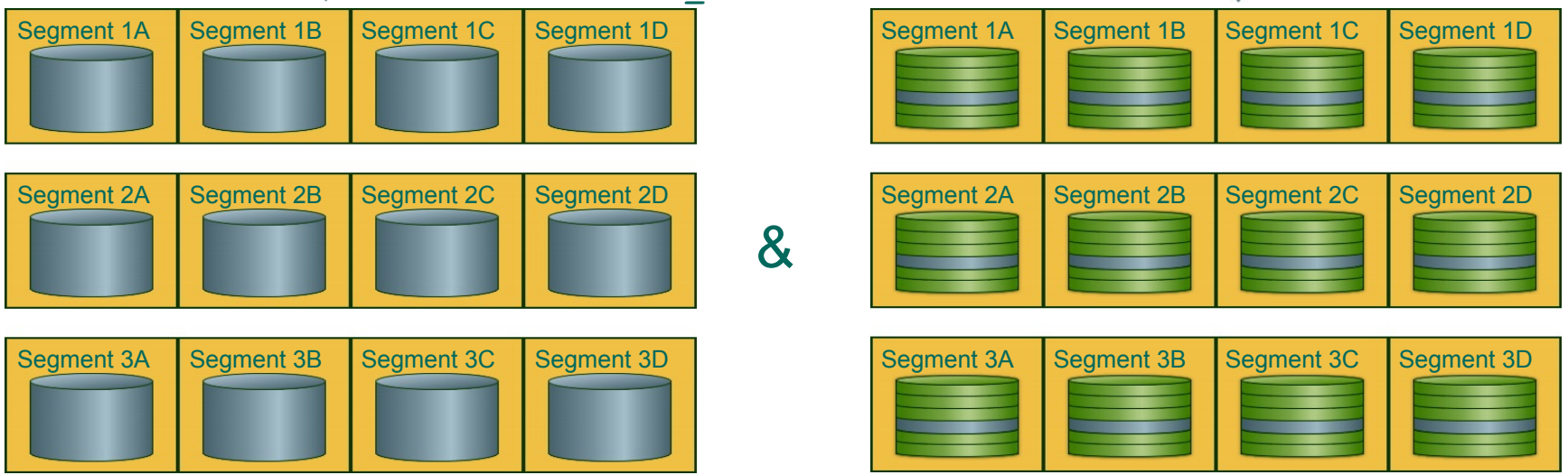
Order		
Order #	Order Date	Customer ID
43	Oct 20 2005	12
64	Oct 20 2005	111
45	Oct 20 2005	42
46	Oct 20 2005	64
77	Oct 20 2005	32
48	Oct 20 2005	12
50	Oct 20 2005	34
56	Oct 20 2005	213
63	Oct 20 2005	15
44	Oct 20 2005	102
53	Oct 20 2005	82
55	Oct 20 2005	55



Pivotal

分布和分区

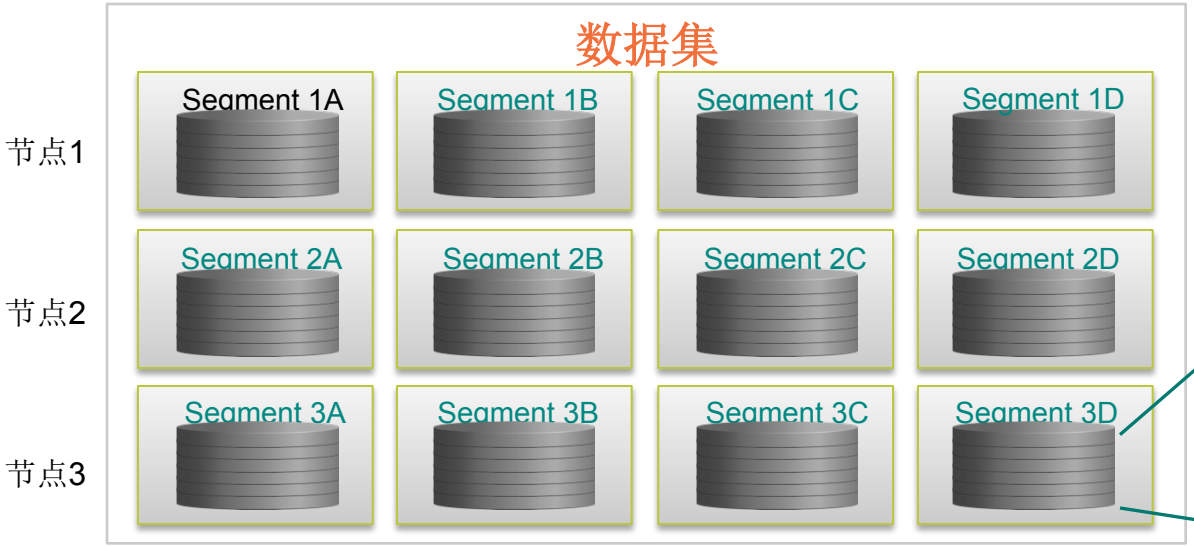
```
SELECT COUNT(*)
  FROM orders
 WHERE order_date >= 'Oct 20 2007'
   AND order_date <  'Oct 27 2007'
```



orders 表数据均匀分布于各个节点

仅仅扫描 orders 表相关的分区

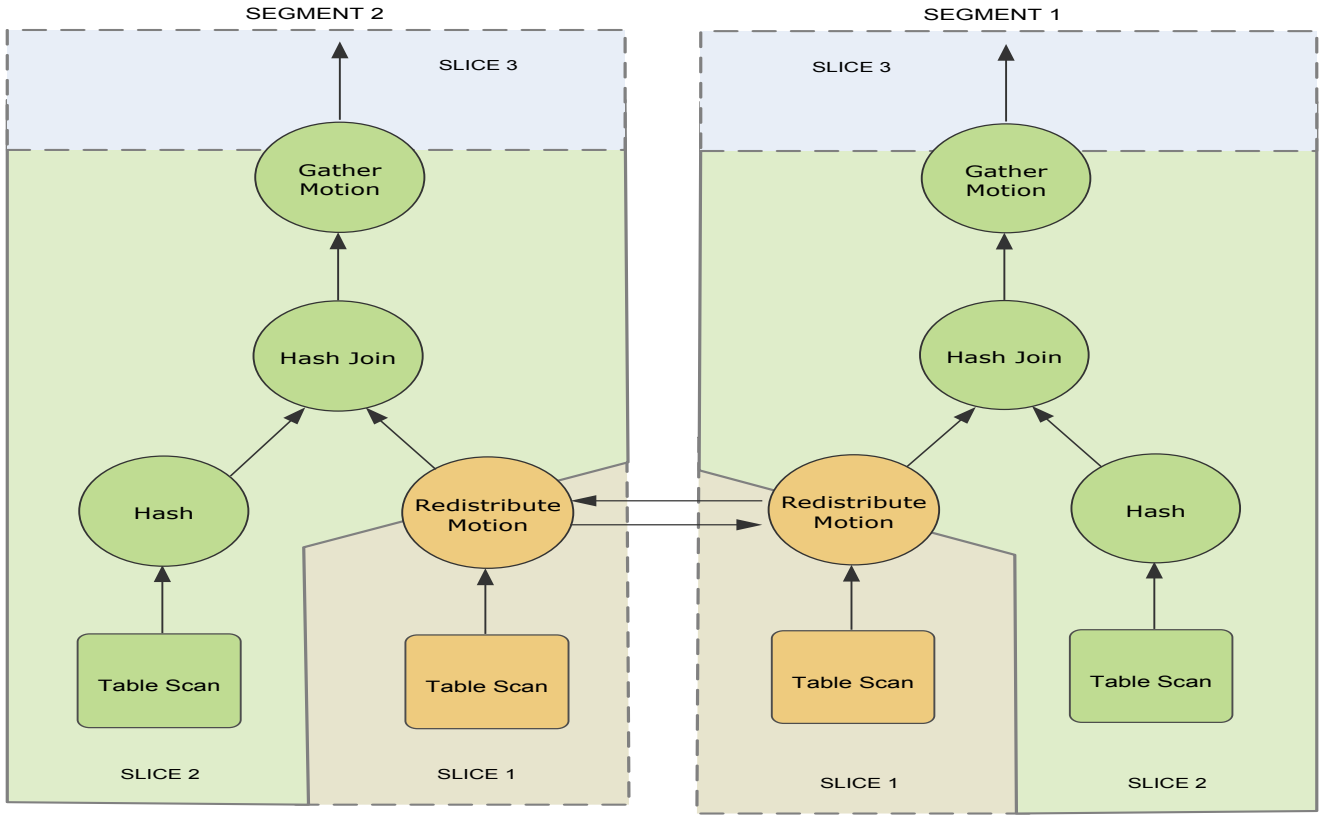
多级分区存储



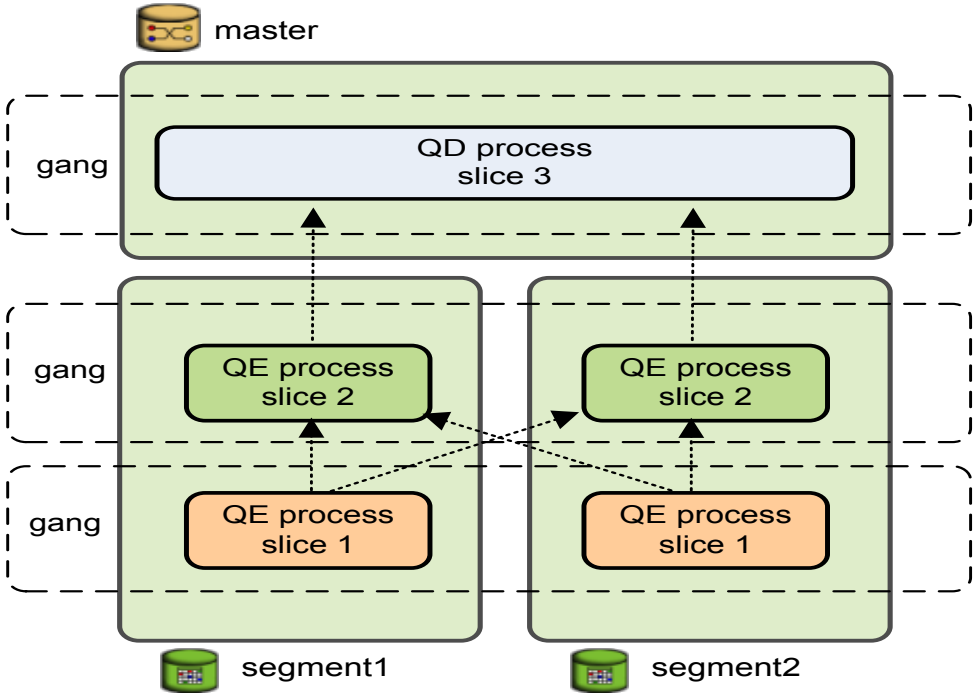
- 哈希Distribution：数据均匀的分布到各个数据节点
- 范围分区： 数据节点内部，根据多种规则分区，降低扫描量

并行查询计划

```
SELECT customer,
       amount
FROM   sales
JOIN   customer
USING (cust_id)
WHERE  date=2008;
```



并行计划的执行


master
gang
QD process
slice 3
gang
QE process
slice 2
QE process
slice 2
gang
QE process
slice 1
QE process
slice 1
segment1
segment2

多态存储

用户自定义数据存储格式

行存储

- 访问多列时速度快
- 支持高效更新和删除
- AO 主要为插入而优化

列存储

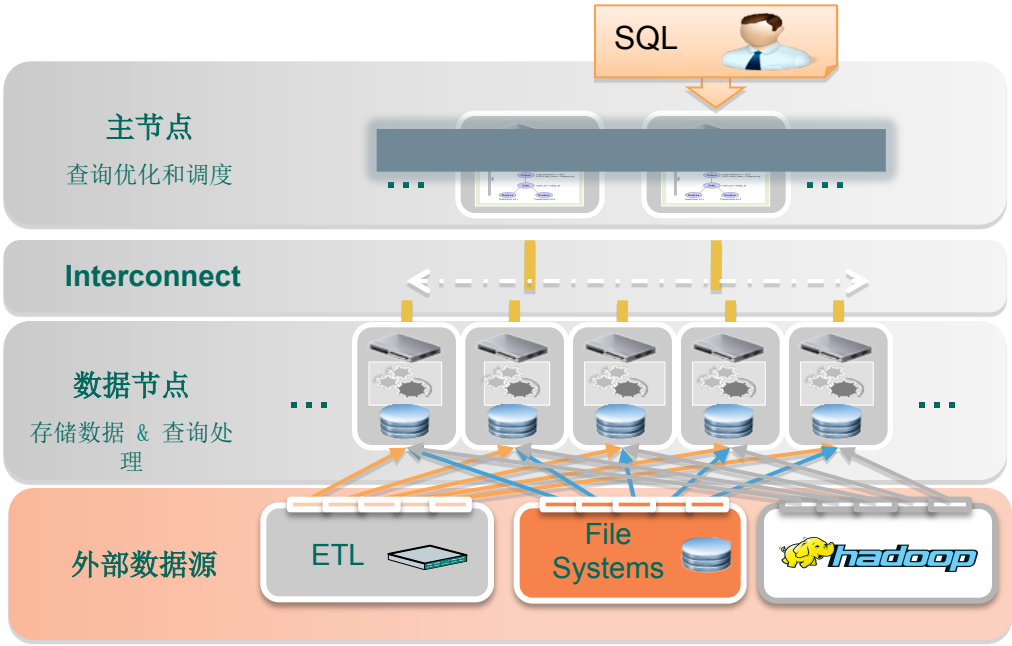
- 列存储更适合压缩
- 查询列子集时速度快
- 不同列可以使用不同压缩方式：gzip (1-9)，quicklz，delta，RLE

外部表

- 历史数据和不常访问的数据存储在 HDFS 或者其他外部系统中
- 无缝查询所有数据
- Text，CSV，Binary，Avro，Parquet 格式

大规模并行数据加载

- 高速数据导入和导出
  - 主节点不是瓶颈
  - 10+ TB/小时/Rack
  - 线性扩展
- 低延迟
  - 加载后立刻可用
  - 不需要中间存储
  - 不需要额外数据处理
- 导入/导出 到&从:
  - 文件系统
  - 任意 ETL 产品
  - Hadoop 发行版

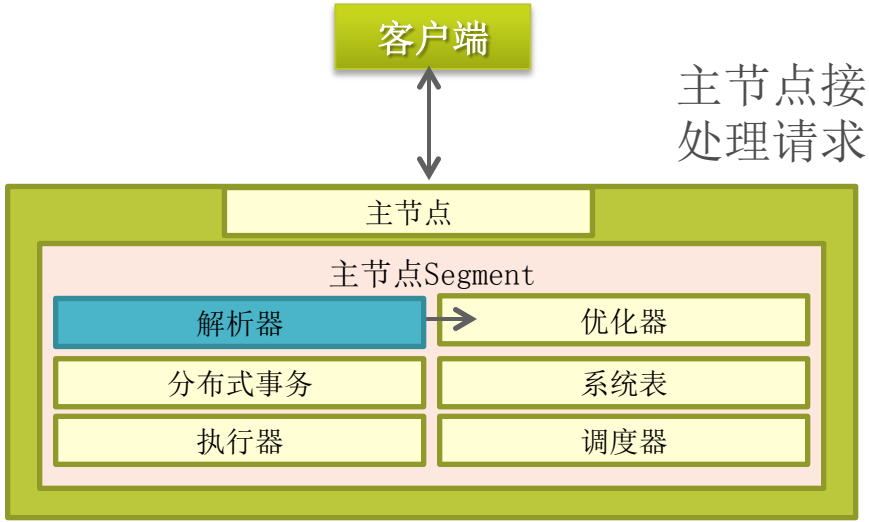


Pivotal

Greenplum 组件

Pivotal

解析器

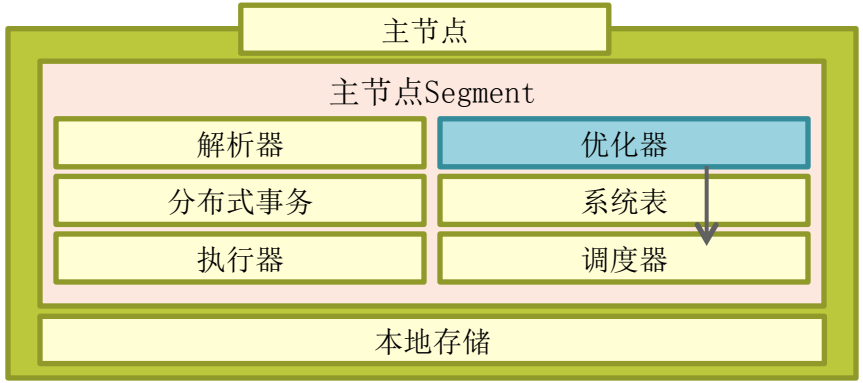


主节点接受客户连接，处理请求，执行认证

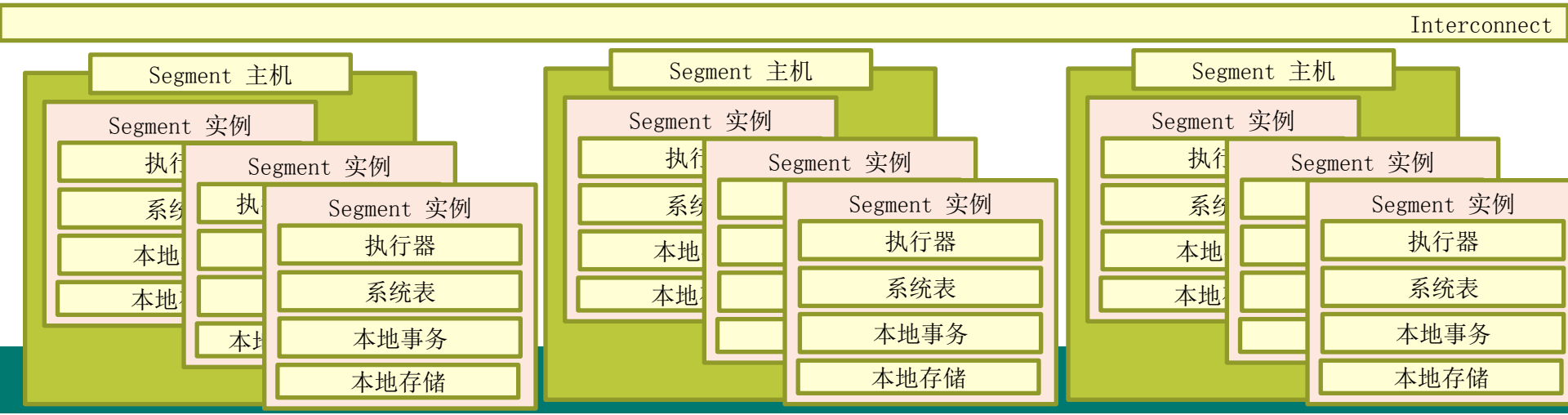
解析器执行词法分析、语法分析并生成 解析树

优化器

处理解析树，生成查询计划

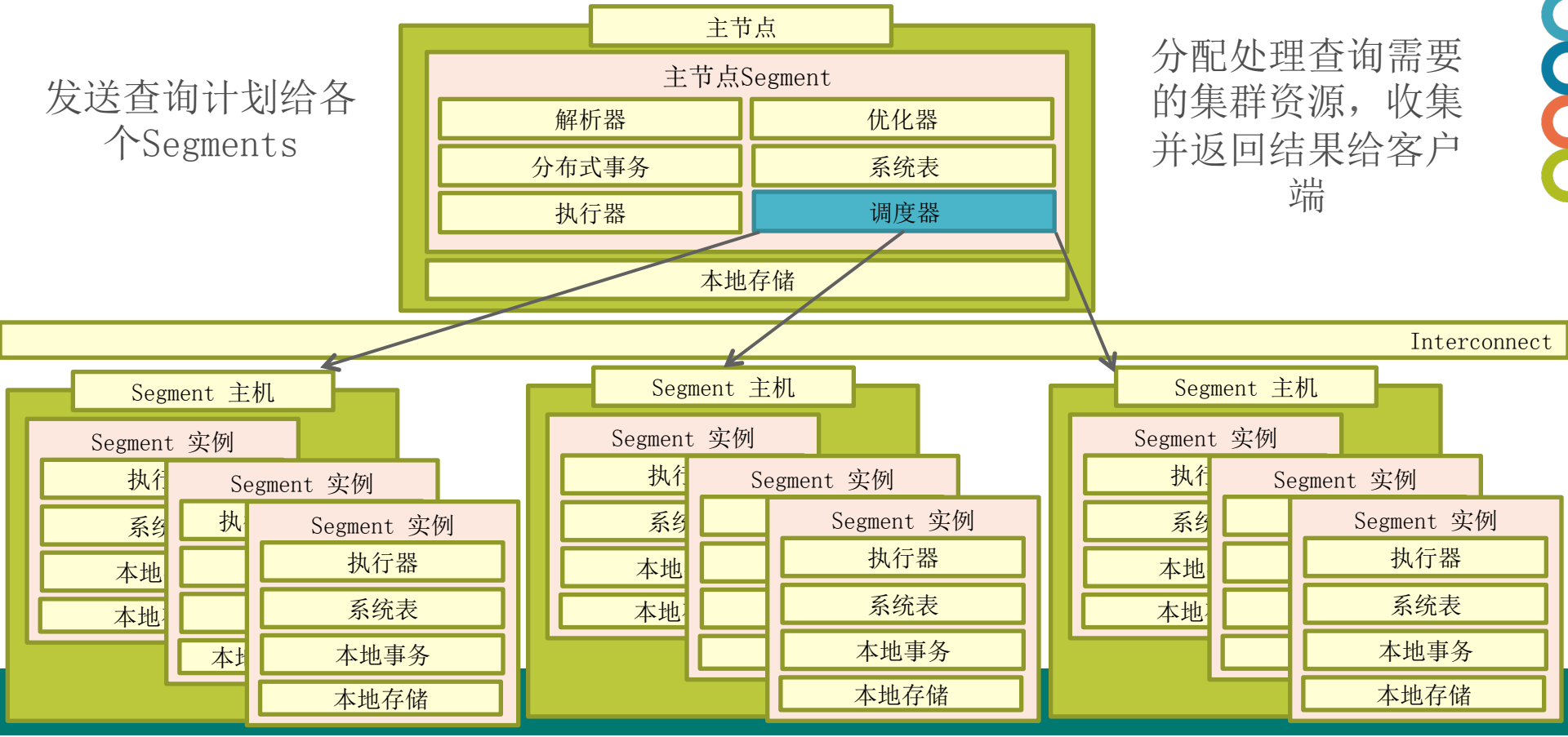


查询计划描述了如何执行查询



调度器

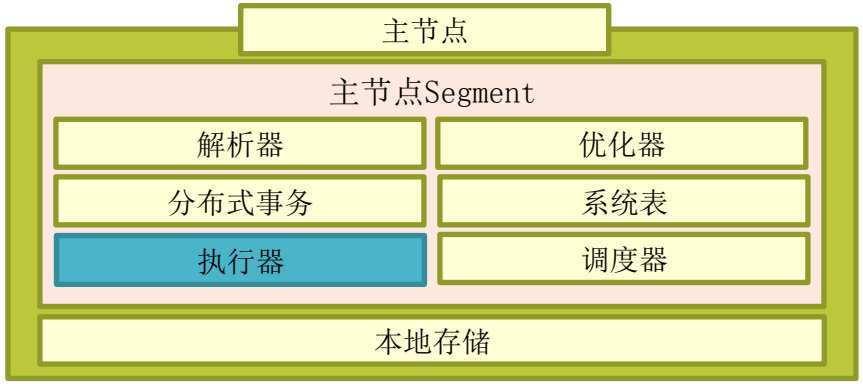
发送查询计划给各个Segments



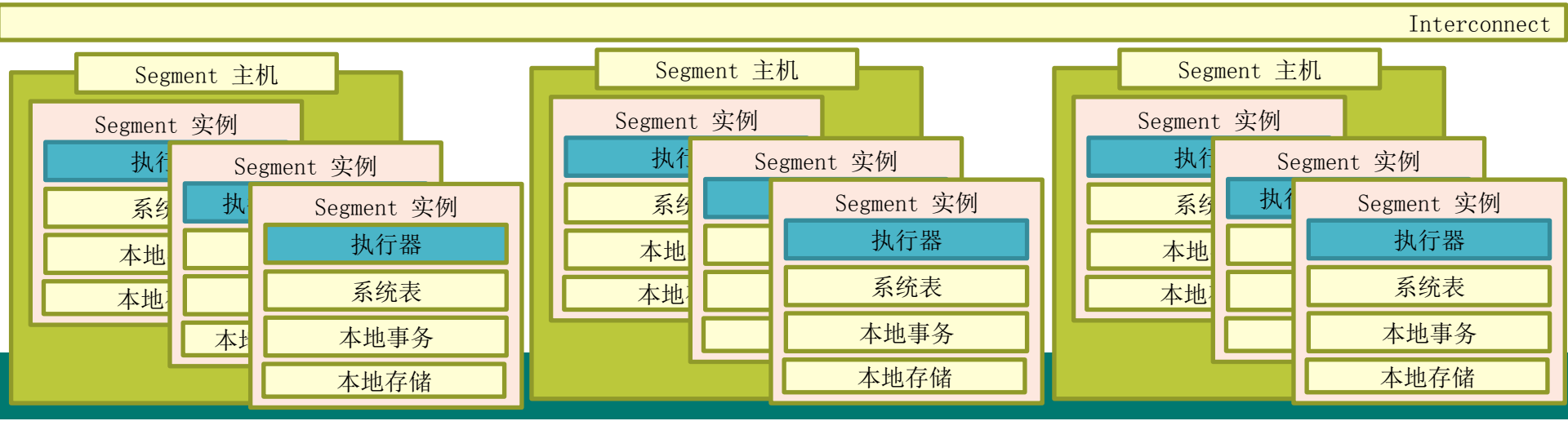
分配处理查询需要的集群资源，收集并返回结果给客户端

执行器

发送查询计划给各个Segments

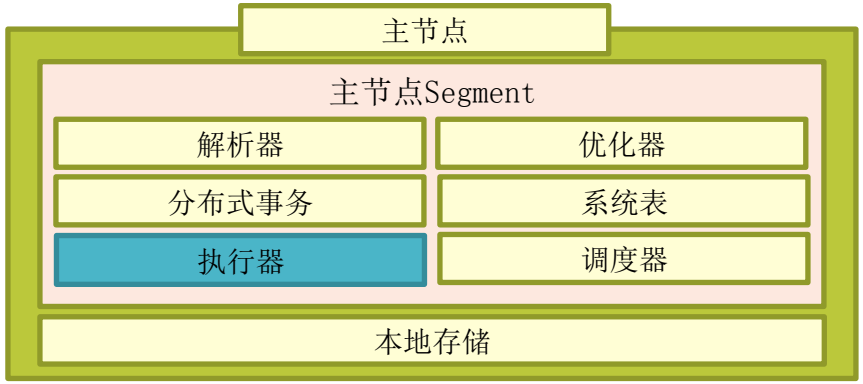


分配处理查询需要的集群资源，收集并返回结果给客户端

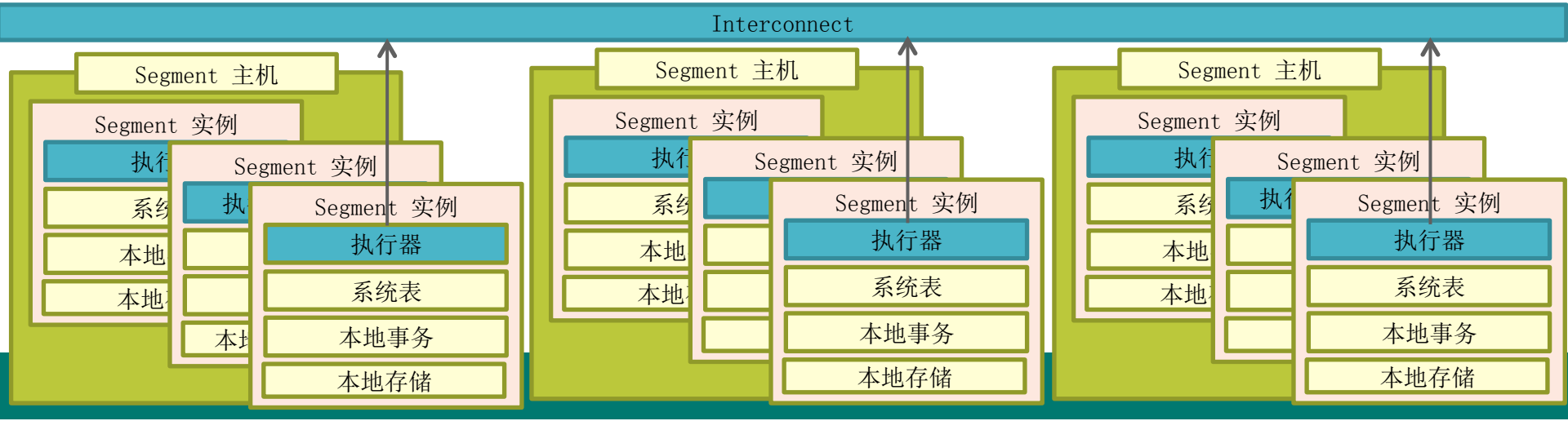


Interconnect

发送查询计划给各个Segments

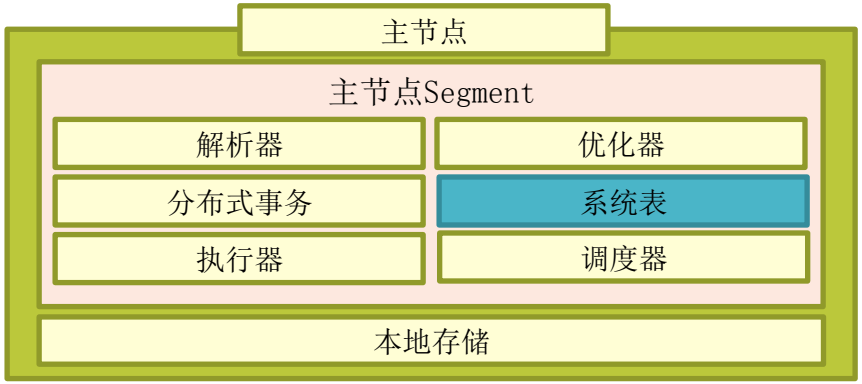


分配处理查询需要的集群资源，收集并返回结果给客户端

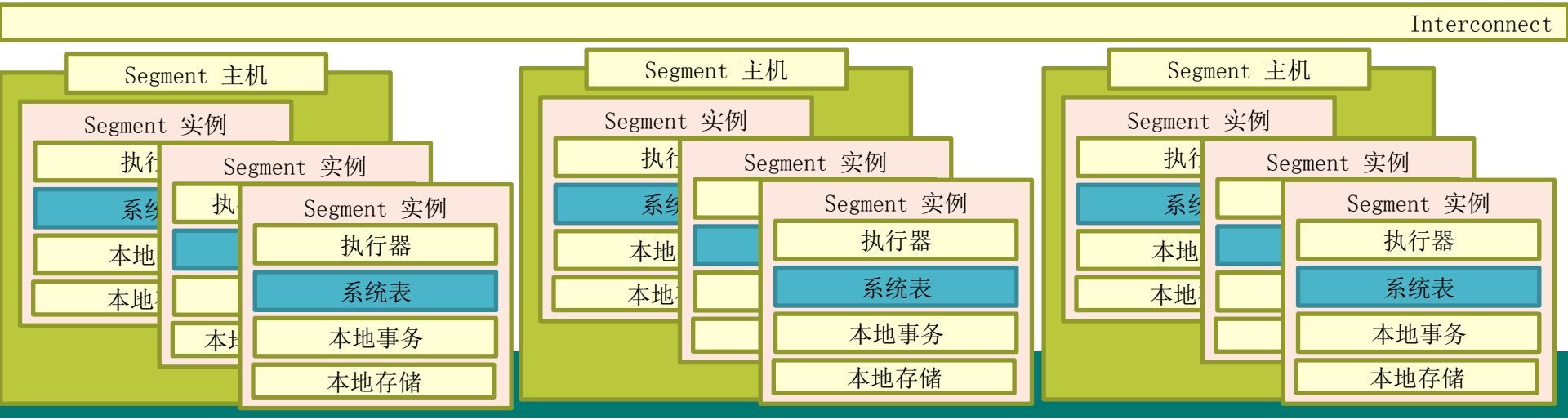


系统表

存储和管理数据库、表、字段的元数据

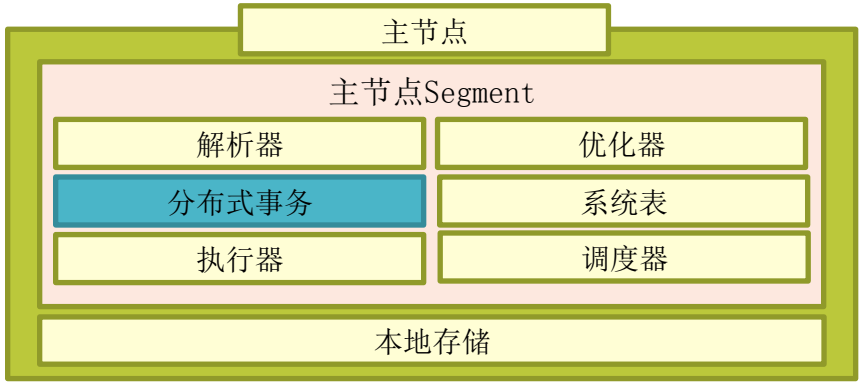


每个节点保存一个拷贝

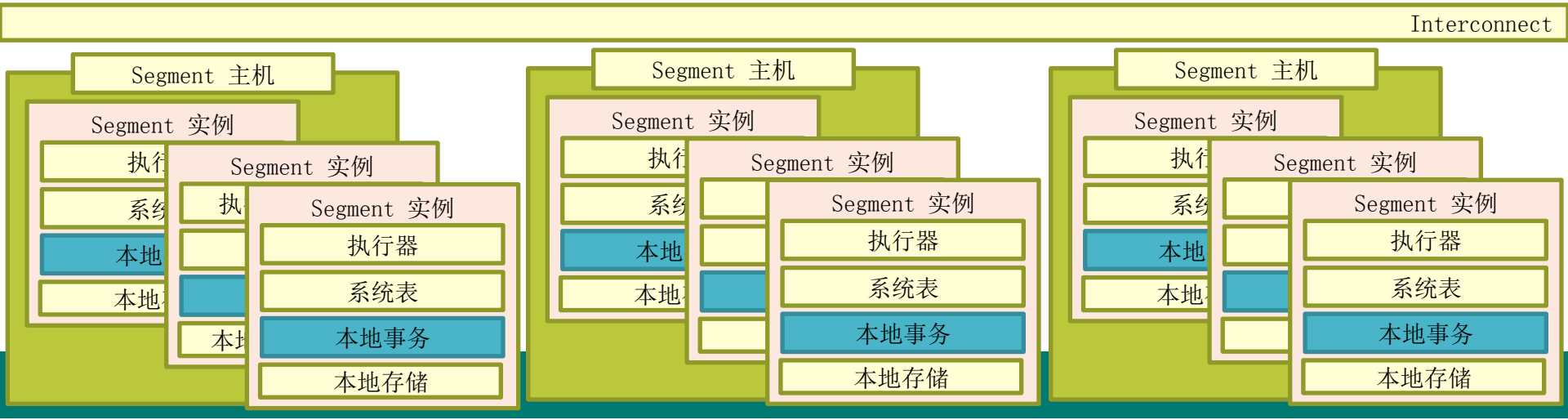


分布式事务

主节点上的分布式事务管理器协调Segment上的提交和回滚操作



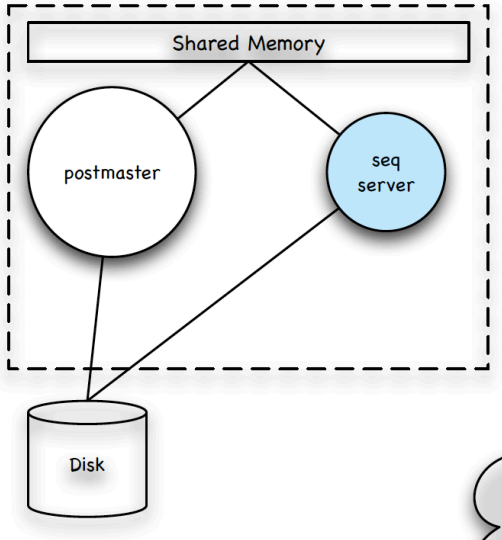
Segments 有自己的事务日志，确定合适提交或回滚自己的事务

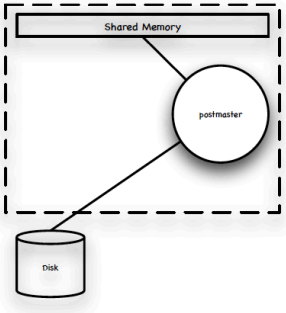


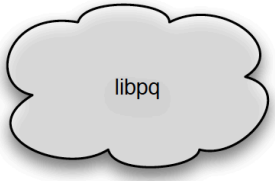
Greenplum 数据库SQL执行流程

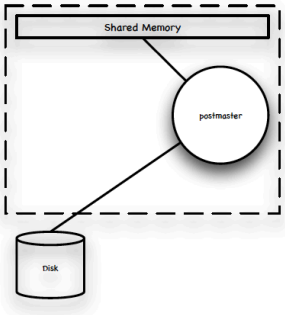
Pivotal

0. The system at rest.

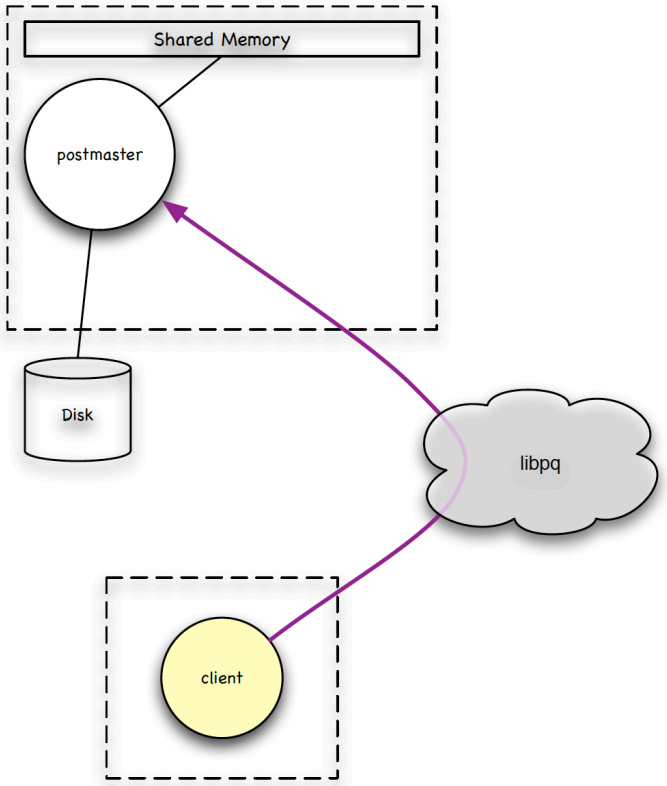


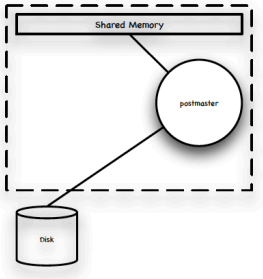


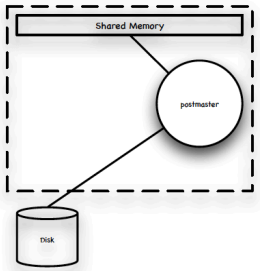




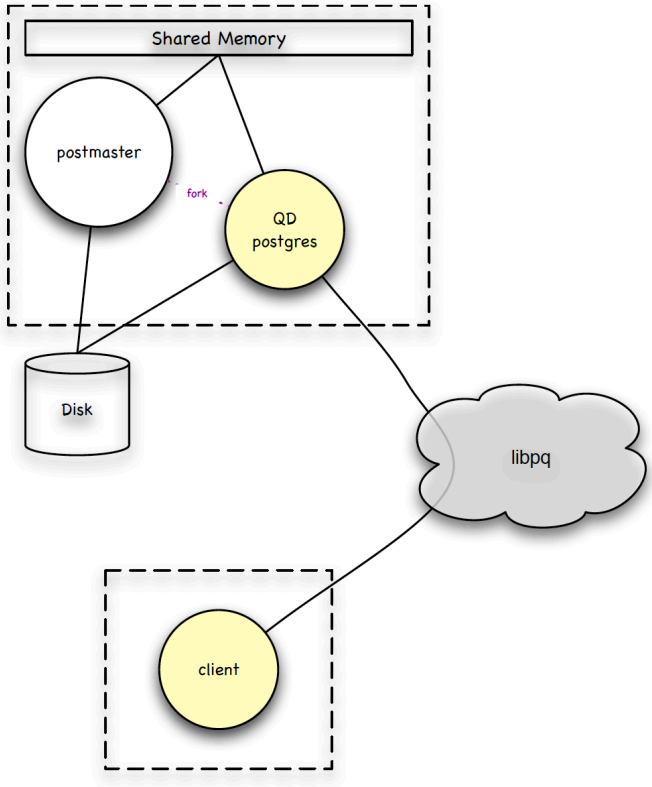
1. Client connects via the entry postmaster.

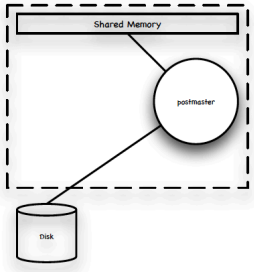


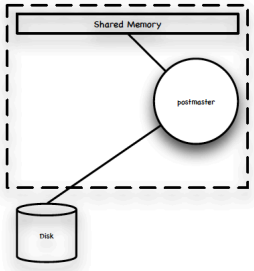




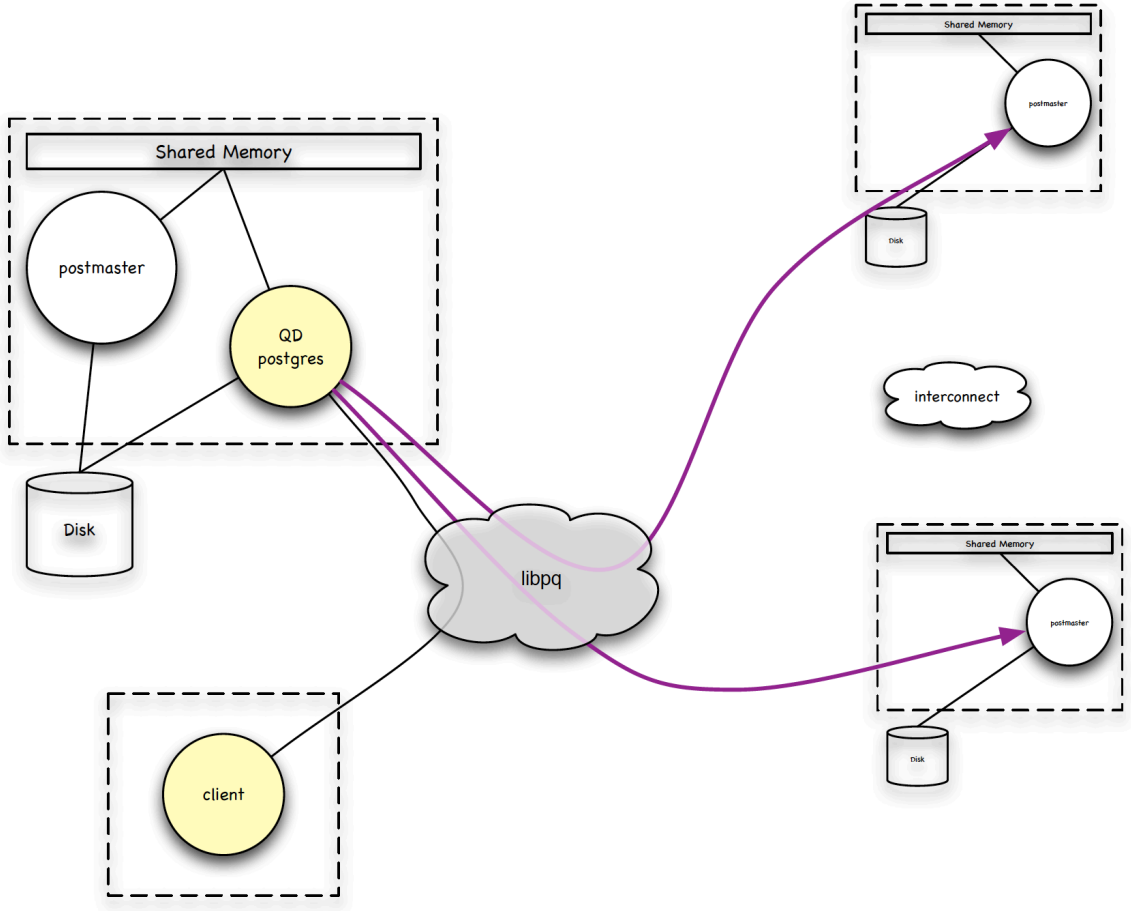
2. Entry postmaster forks a new backend -- the QD.



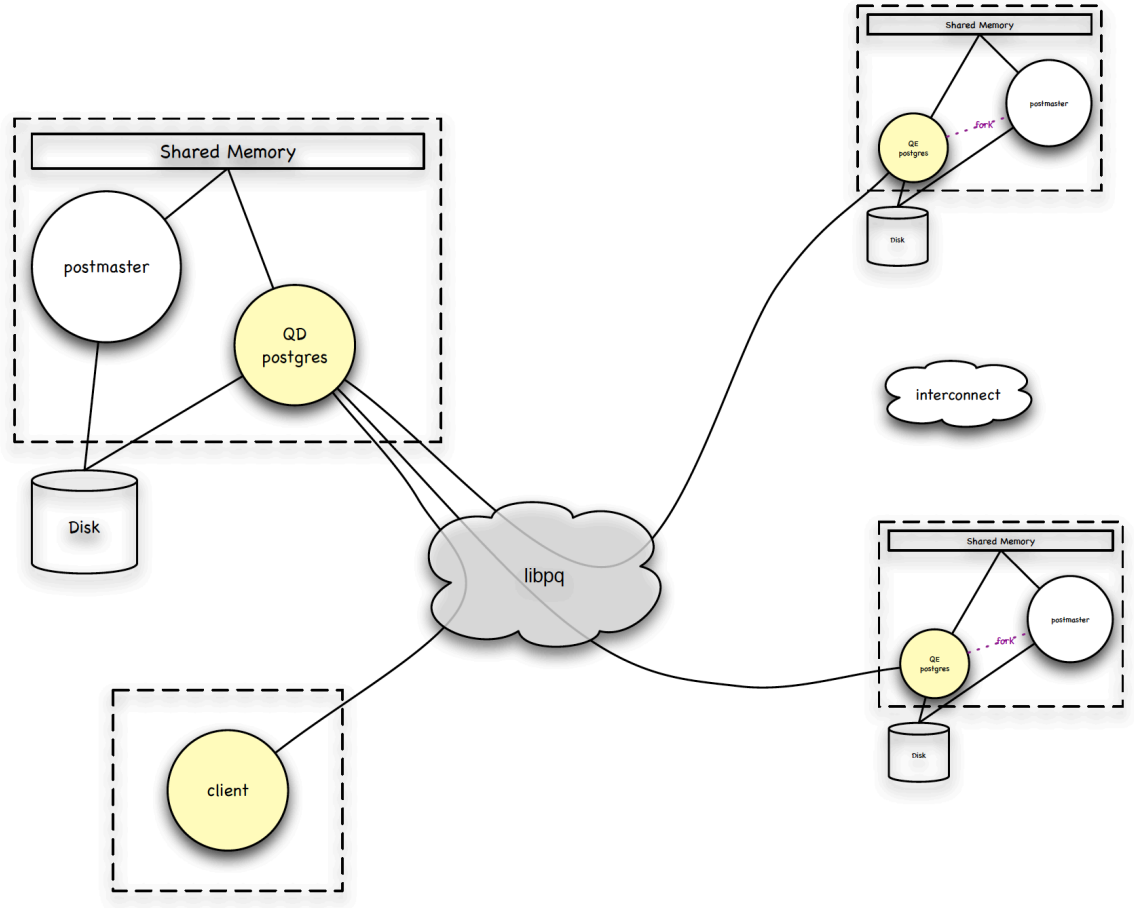




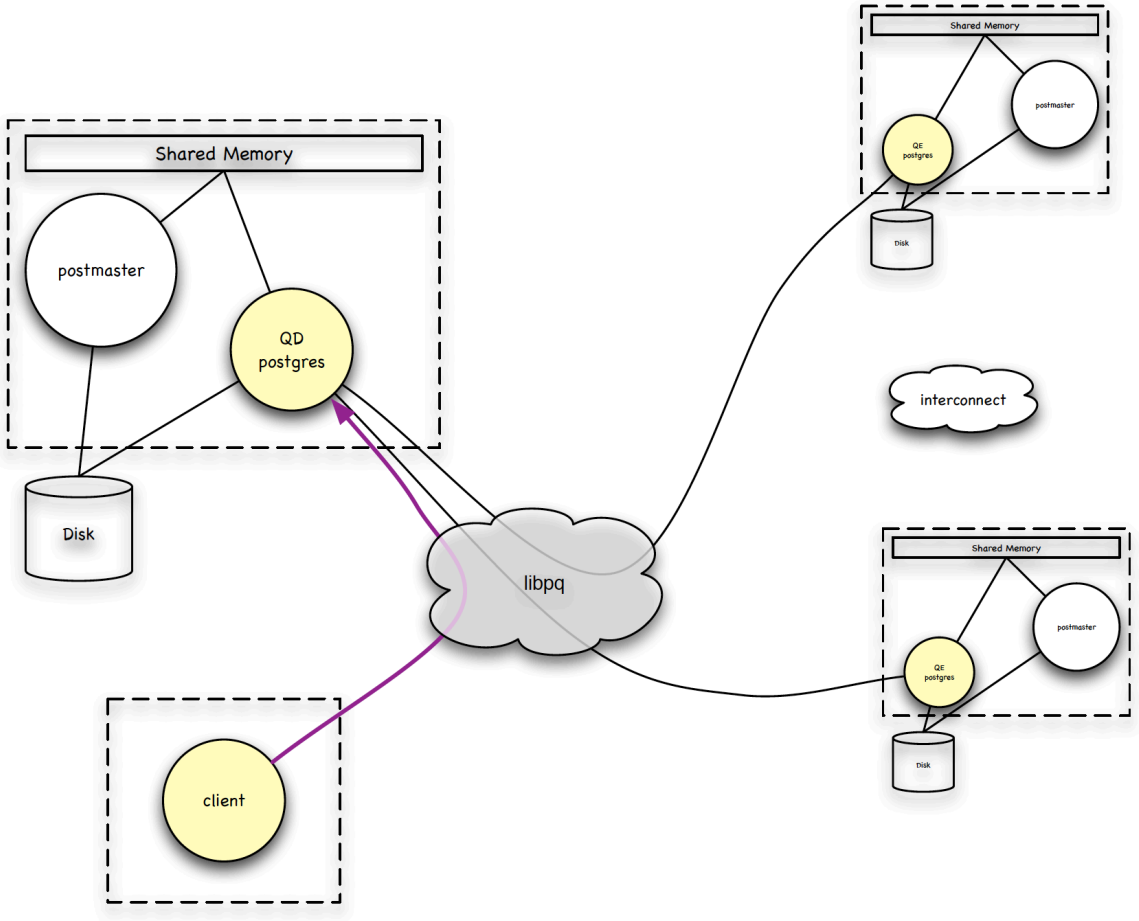
3. QD connects to segments via the segment postmasters.



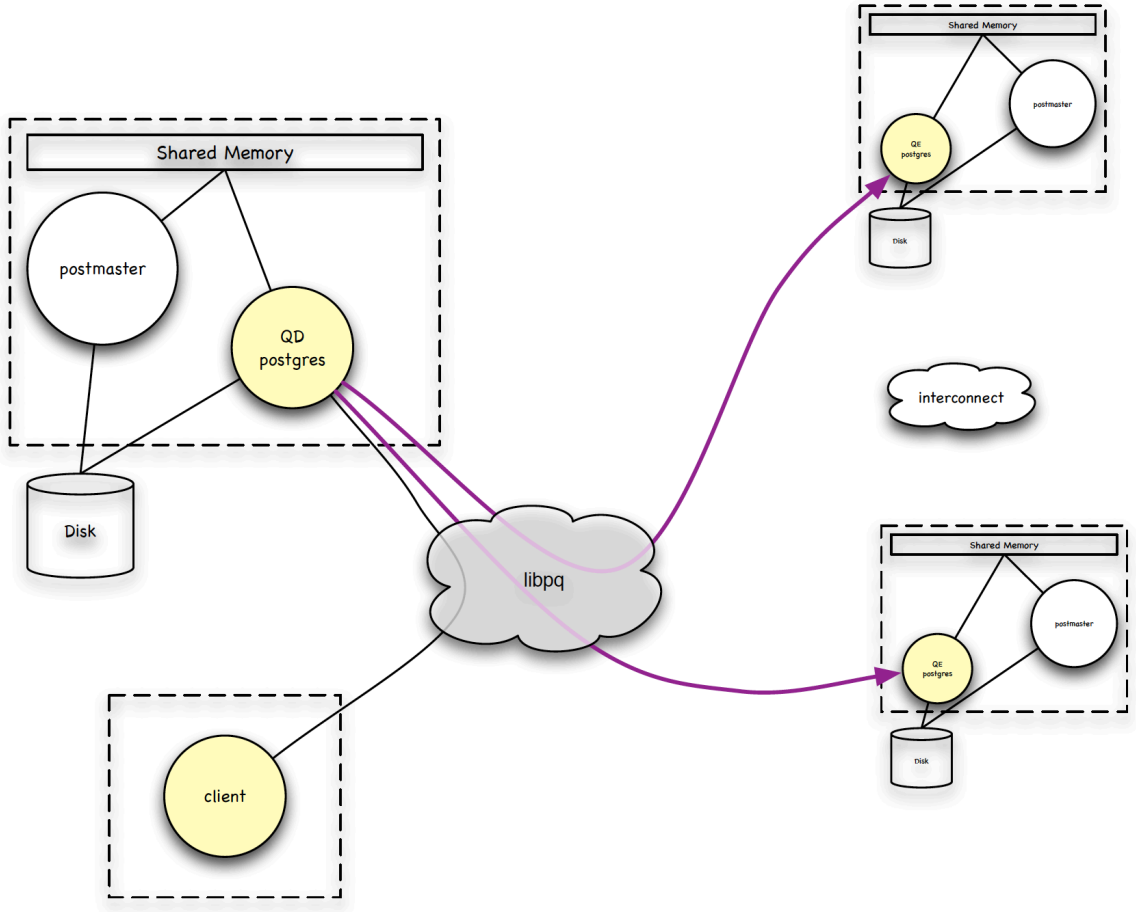
4. Segment postmasters fork initial gang of QEs.



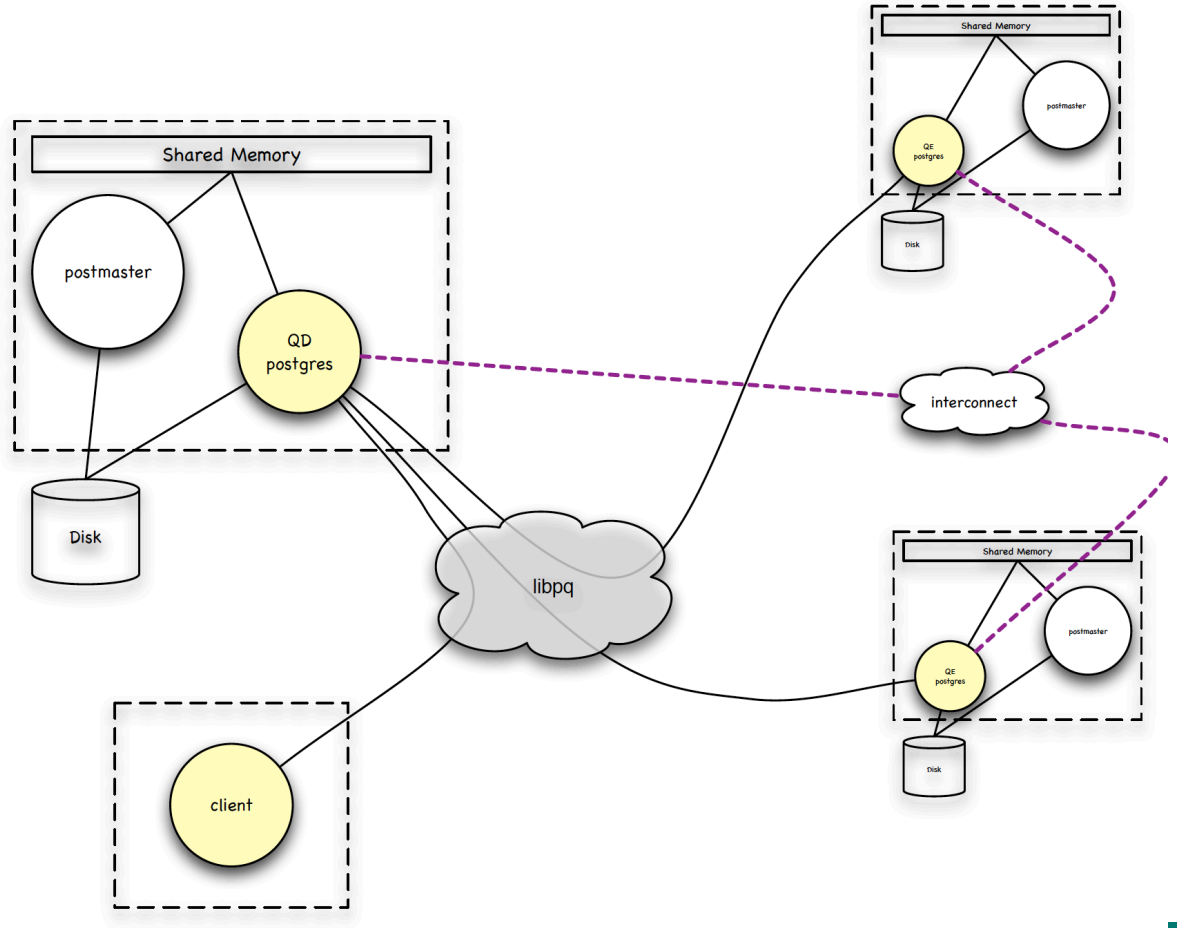
5. Client submits a query to the QD.



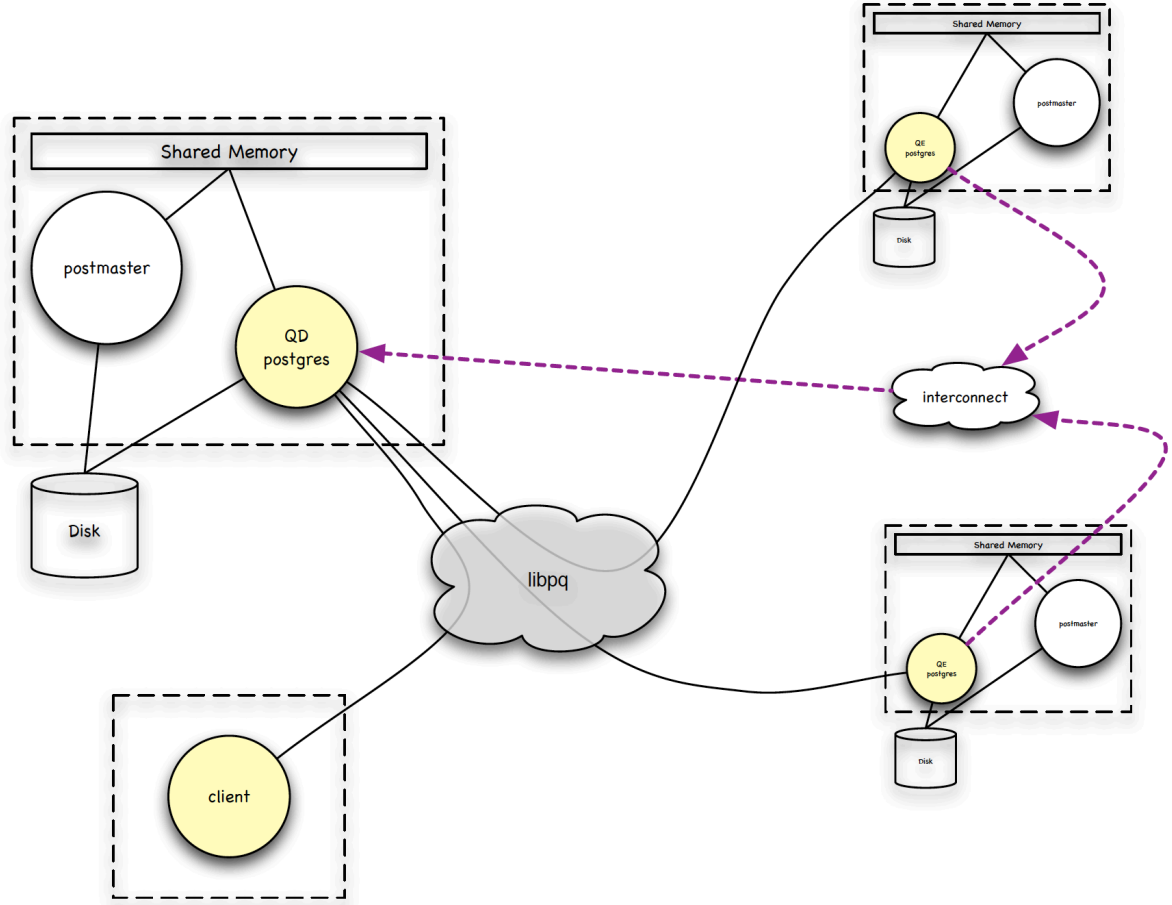
6. QD plans query and submits plans to QEs.



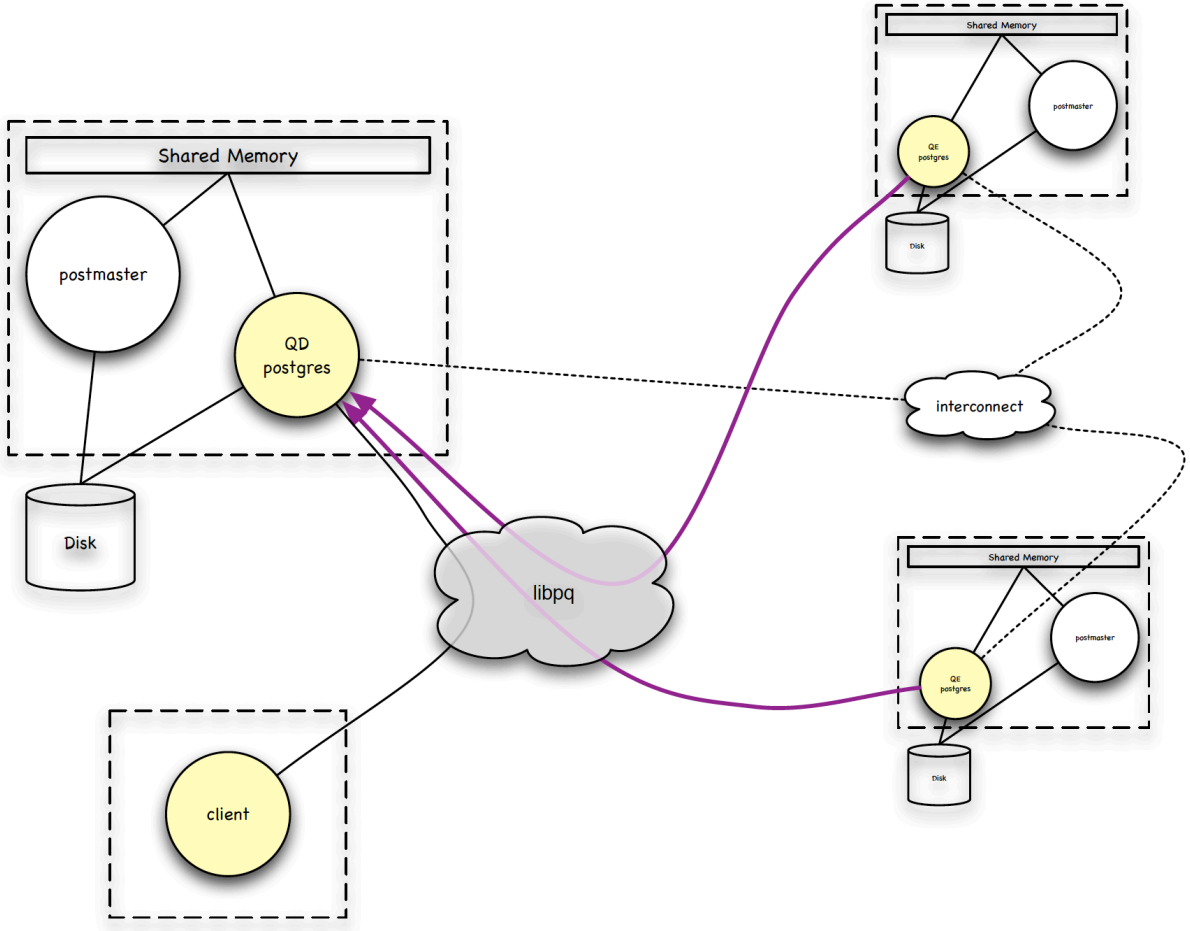
7. QD and QEs setup interconnect routes according to plan.



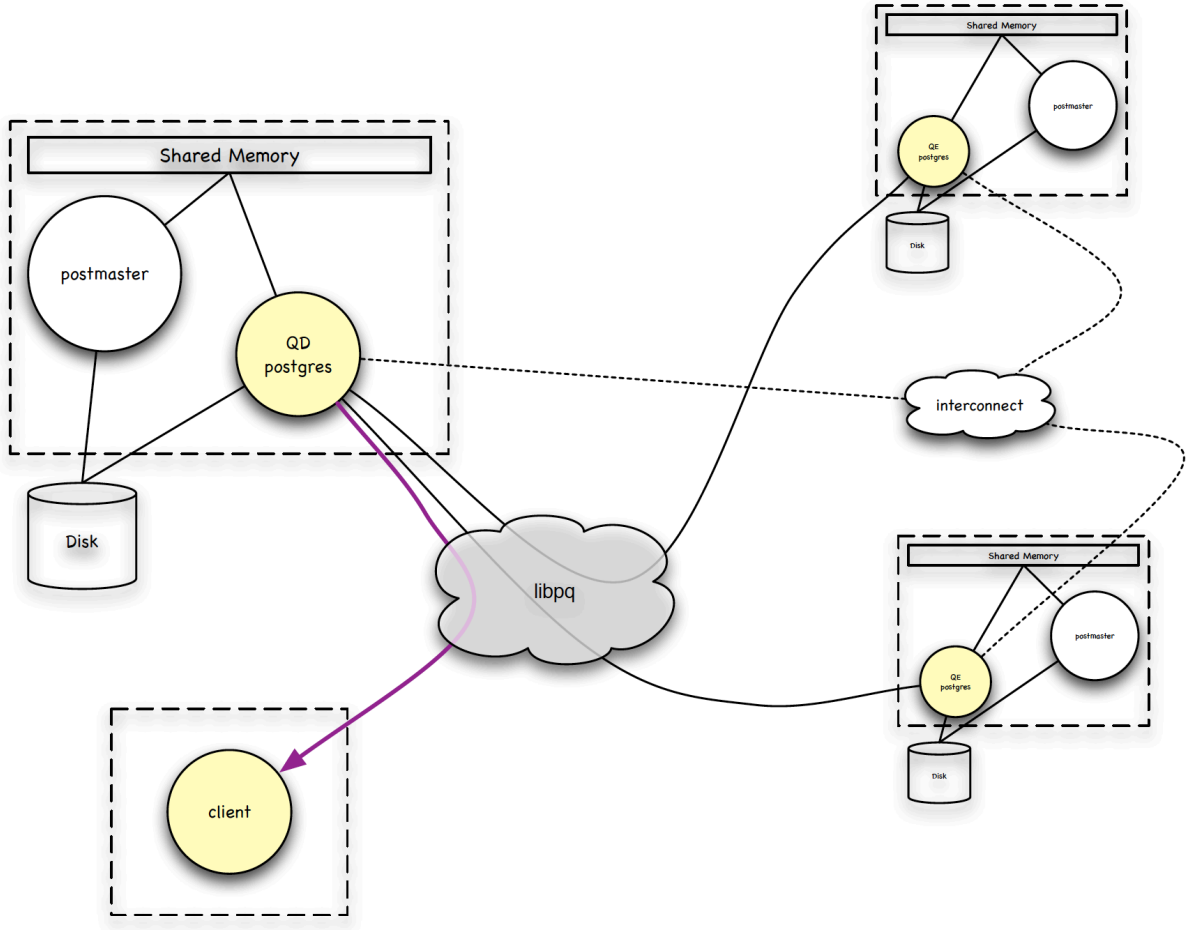
8. QD and QEs execute their slices sending tuples up the slice tree.



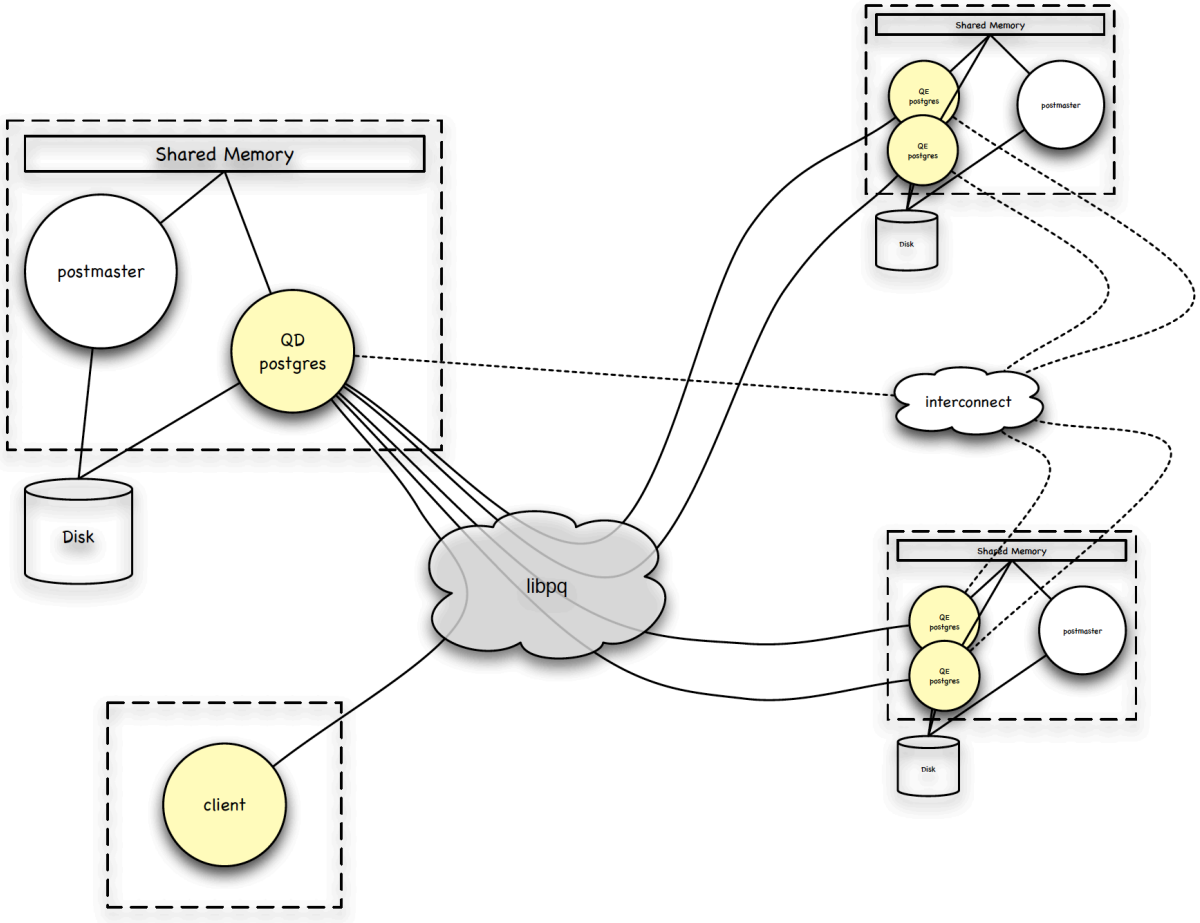
9. QEs return status to QD.



10. QD returns result set and status to the client.



Note that additional gangs of QEs are added as required by plans.



Greenplum 开源

Pivotal

最先进的开源分布式数据库

- 2015/03 宣布开源计划，2015/10/27 正式开源
- Apache License
- http://greeplum.org

相关链接

- 主站点：http://greenplum.org/
- 源代码：https://github.com/greenplum-db/gpdb
- 邮件列表：http://greenplum.org/#mailing-lists
- 如何贡献：http://greenplum.org/#contribute
- 活动：http://greenplum.org/#events
  - GPDB Meetup: 2016/01/12 北京
    http://www.meetup.com/Greenplum-Community/events/226900679/?a=co2_grp&gj=co2&rv=co2

中文社区

- 欢迎大学、研究机构、公司和客户参与到社区中，共建世界一流的大数据数据库！
- 欢迎和Pivotal中国研发中心合作
- 微信公众号：Greenplum(即将上线)
- 微信群：
- http://gpdb.rocks
- QQ 群：99194625

Pivotal

热招职位：

- Greenplum 数据库内核开发（C, Python）
- Greenplum 文本检索产品开发（C/C++, Python, Java）
- Greenplum GIS 产品开发（C）
- Greenplum 管理控制台前端和后端开发（golang, HTML5）

欢迎参与到世界一流的数据库内核团队

Q&A

Appendix

Pivotal

GPDB 高可用性

- 主节点（Master）高可用
  - Warm Standby
    - 主节点系统表副本
  - 避免单点故障
  - 主节点和从主节点间同步复制
    - 使用流复制
- 数据节点（Segment）高可用
  - 每个Segment都配备一个Mirror
    - 使用文件块级别的复制
  - 自动故障切换（failover）

故障检测和恢复

- ftsprobe 故障检测进程使用心跳检测segments是否发生故障
- gpstate 工具查看primary和mirror segments 的状态
- 查询 *gp_segment_configuration* 系统表可以查看故障Segment的详细信息
  - $ psql -c "SELECT * FROM gp_segment_configuration WHERE status='d';"
- 当 ftsprobe 不能连接到某个 Segment 时标记其为宕机
  - 系统管理员可以使用 gprecoverseg 工具恢复宕机节点
- 自动failover到镜像 Segment
  - 之后的连接切换到镜像节点