

开源大数据引擎:Greenplum架构分析

姚延栋
yyao@pivotal.io
Greenplum 研发总监

Greenplum 简介

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(MPP): 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

Morgan Stanley Awards Pivotal the CTO Award for Innovation

By

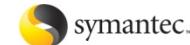
Published: June 5, 2014 11:00 a.m. ET



Aa Aa

SAN FRANCISCO, CALIFORNIA, Jun 05, 2014 (Marketwired via COMTEX) -- Pivotal, the software company at the intersection of big data, PaaS, and agile development, today announced that it has been awarded Morgan Stanley's prestigious "CTO Award for Innovation". Each year, Morgan Stanley awards its 'CTO Award for Innovation' honoring a technology solution that has been deemed innovative and has made a significant impact on Morgan Stanley's business. The award recognizes the role that Pivotal Greenplum plays within the Morgan Stanley's next generation Big Data platforms for financial reporting and risk management, allowing massive scale while lowering the firm's total cost of ownership.

客户



合作伙伴



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
open source business intelligence™

 + a b | e a u
S O F T W A R E

CSC

KARMA SPHERE

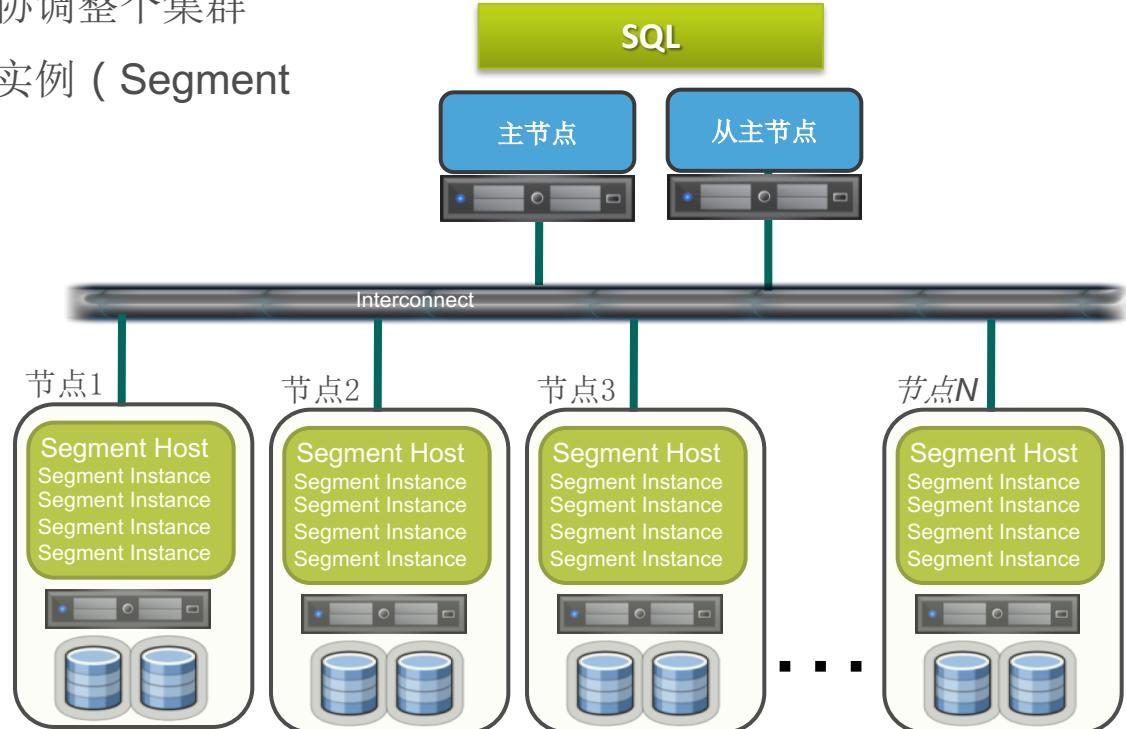
Greenplum 架构

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 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）	语言支持 标准 SQL 支持 支持 MapReduce SQL 2003 OLAP 扩展 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 无共享架构

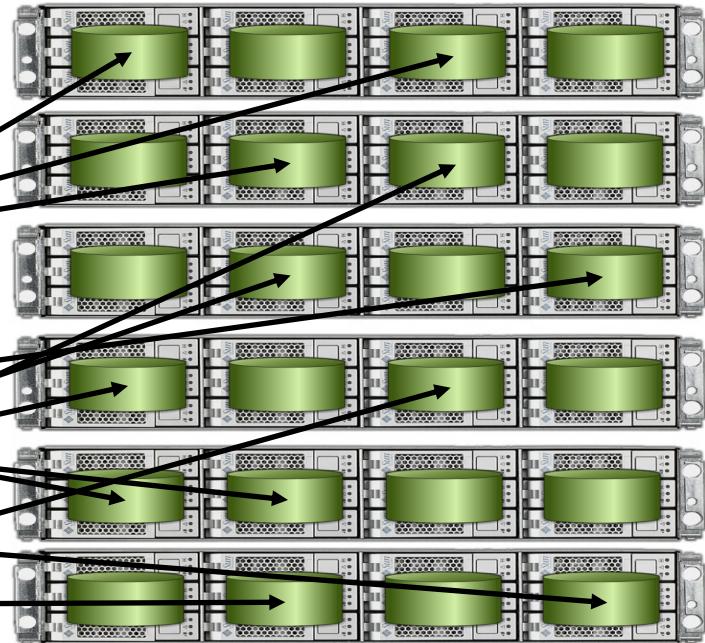
- 主节点和从主节点，主节点负责协调整个集群
- 一个数据节点可以配置多个节点实例（Segment Instances）
- 节点实例并行处理查询（SQL）
- 数据节点有自己的CPU、磁盘和内存（Share nothing）
- 高速Interconnect处理持续数据流（Pipelining）



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



数据节点并行数据访问和处理

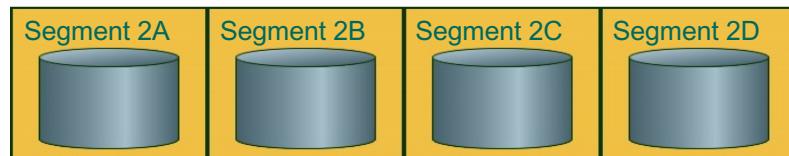
```
SELECT COUNT(*)  
  FROM orders  
 WHERE 4423a328 < 'Oct 20 2007'  
   AND order_date < 'Oct 27 2007'
```



Master



Segment Processors



Each Segment Scans Data Simultaneously in Parallel

分布和分区

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



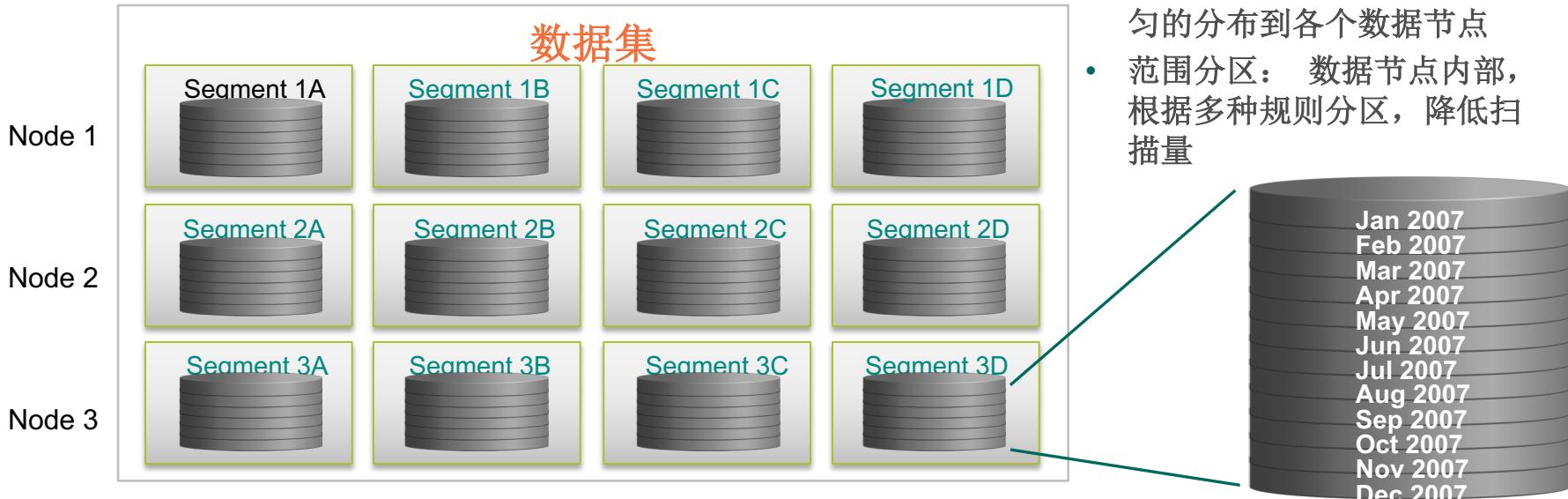
orders 表数据均匀分布于各个节点

&



仅仅扫描 orders 表相关的分区

存储：多级分区



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
- AIO 主要为插入而优化

列存储

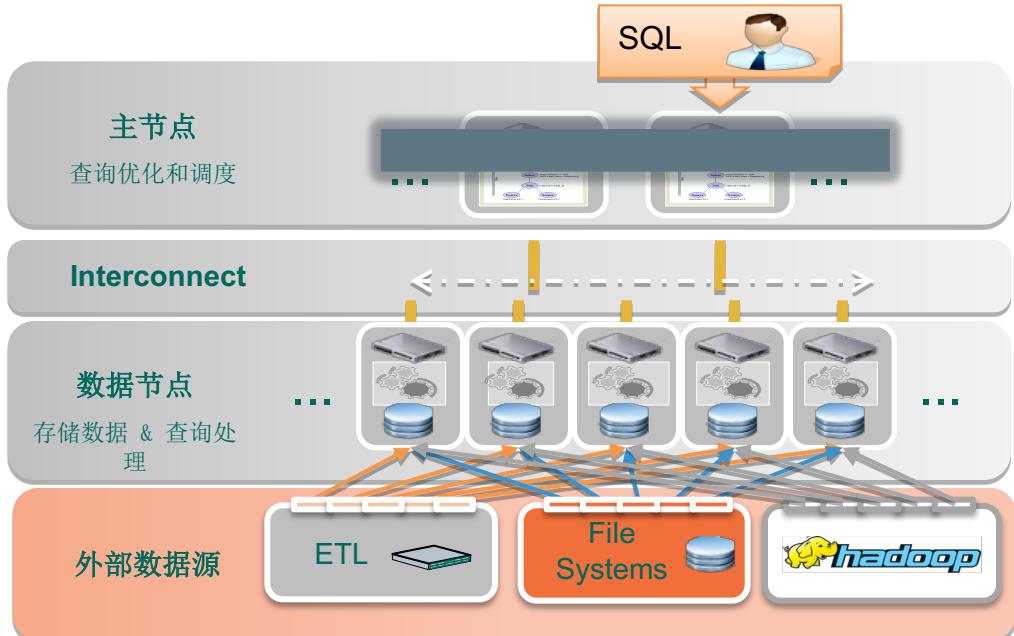
- 列存储更适合压缩
- 查询列子集时速度快
- 不同列可以使用不同压缩方式: gzip (1-9), quicklz, delta, RLE

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

加载：大规模并行数据加载

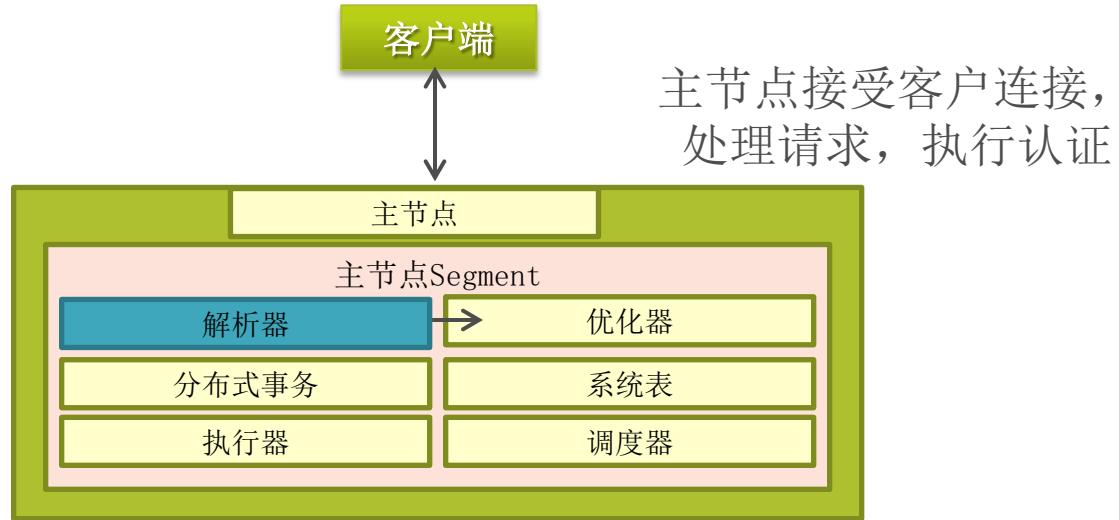
- 高速数据导入和导出
 - 主节点不是瓶颈
 - 10+ TB/小时/Rack
 - 线性扩展
- 低延迟
 - 加载后立刻可用
 - 不需要中间存储
 - 不需要额外数据处理
- 导入/导出 到&从：
 - 文件系统
 - 任意 ETL 产品
 - Hadoop 发行版



Greenplum 组件

解析器

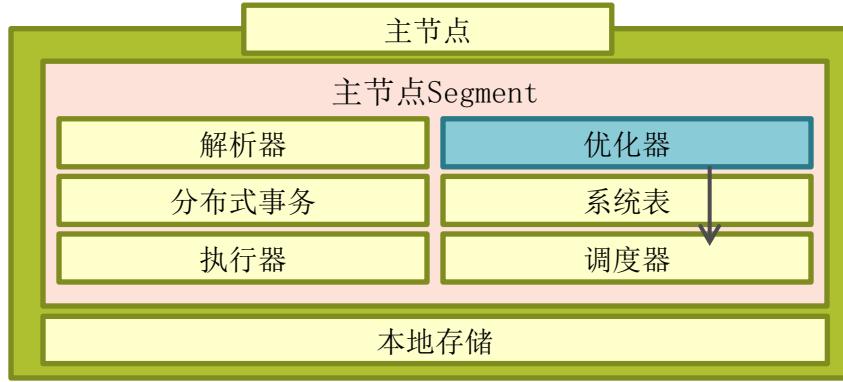
解析器执行词法分析、语法分析并生成 解析树



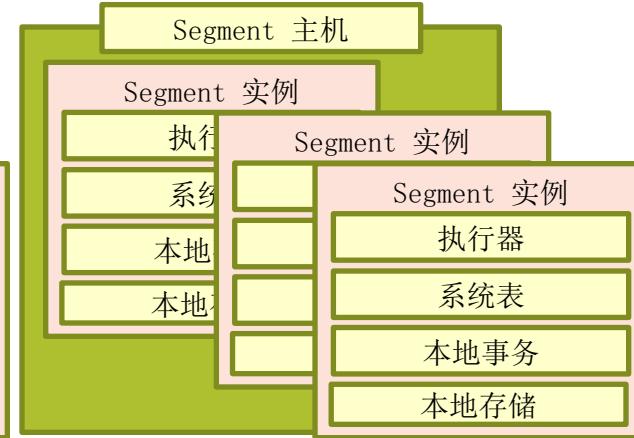
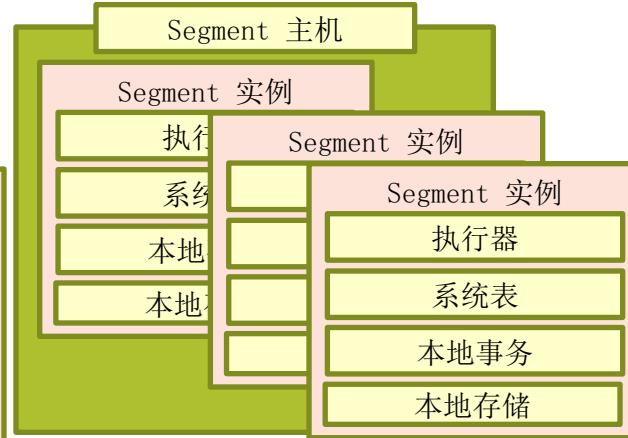
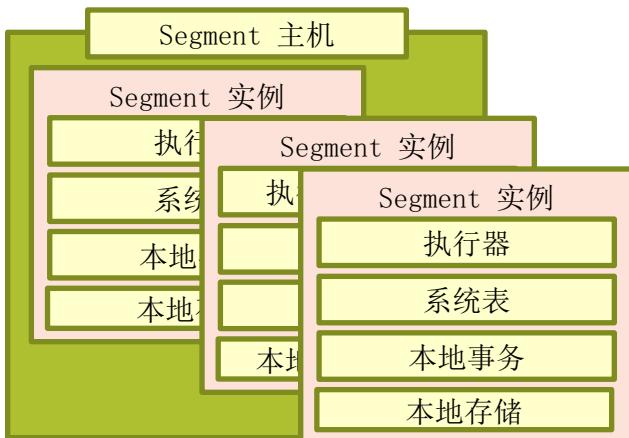
Pivotal 查询优化器

处理解析树，生成
查询计划

查询计划描述了如
何执行查询

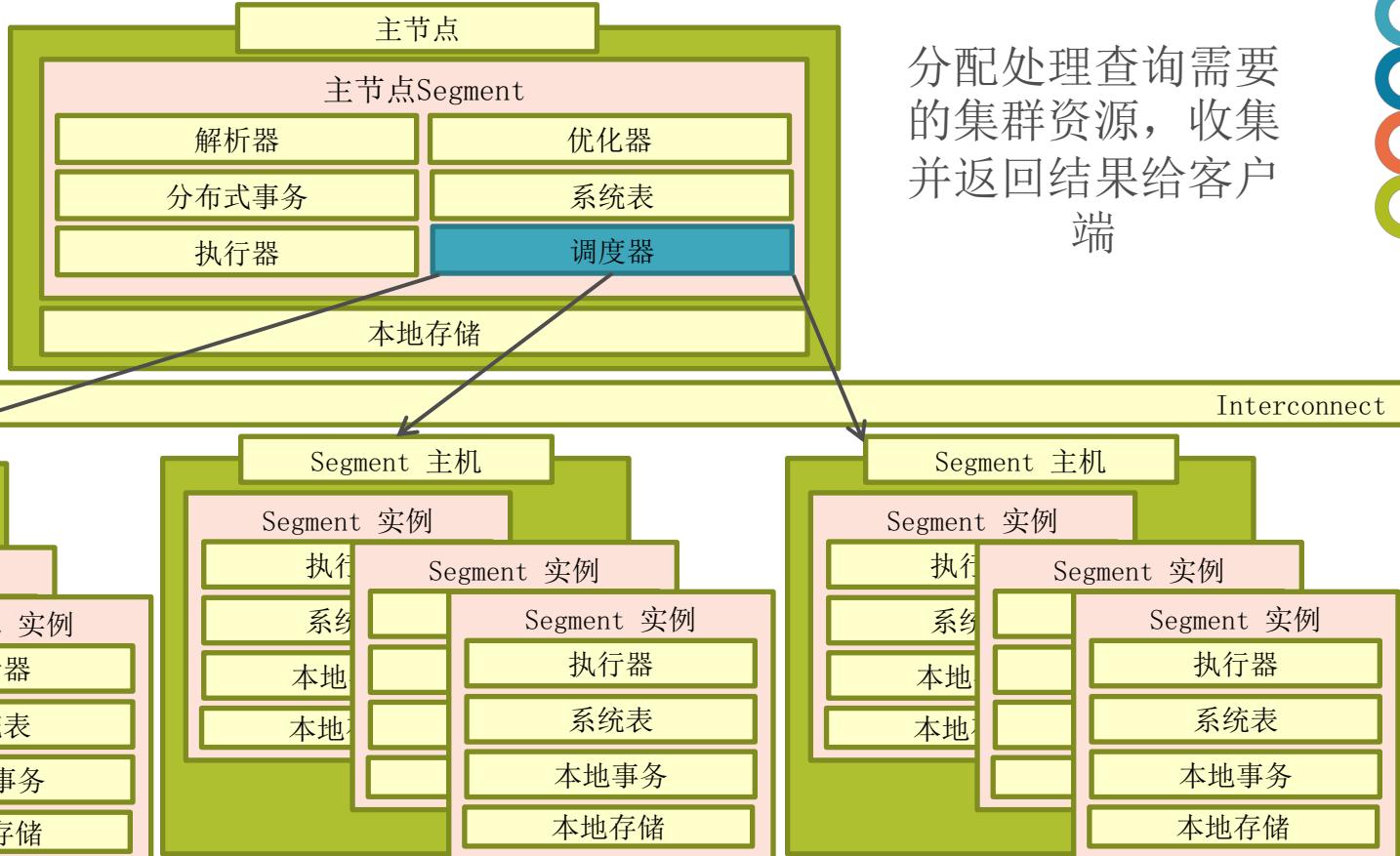


Interconnect



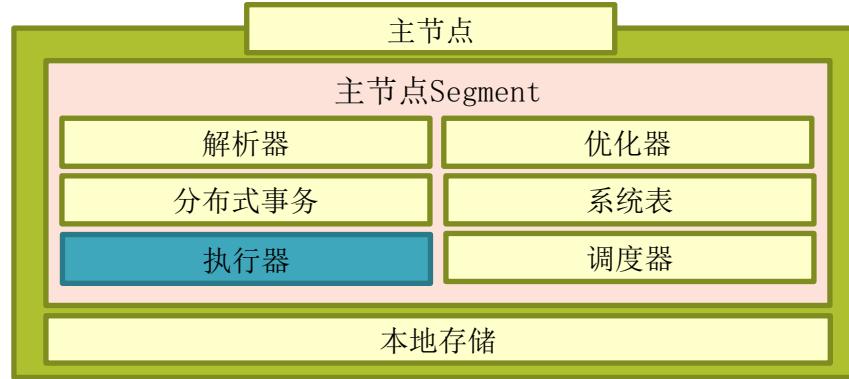
调度器

发送查询计划给各个Segments



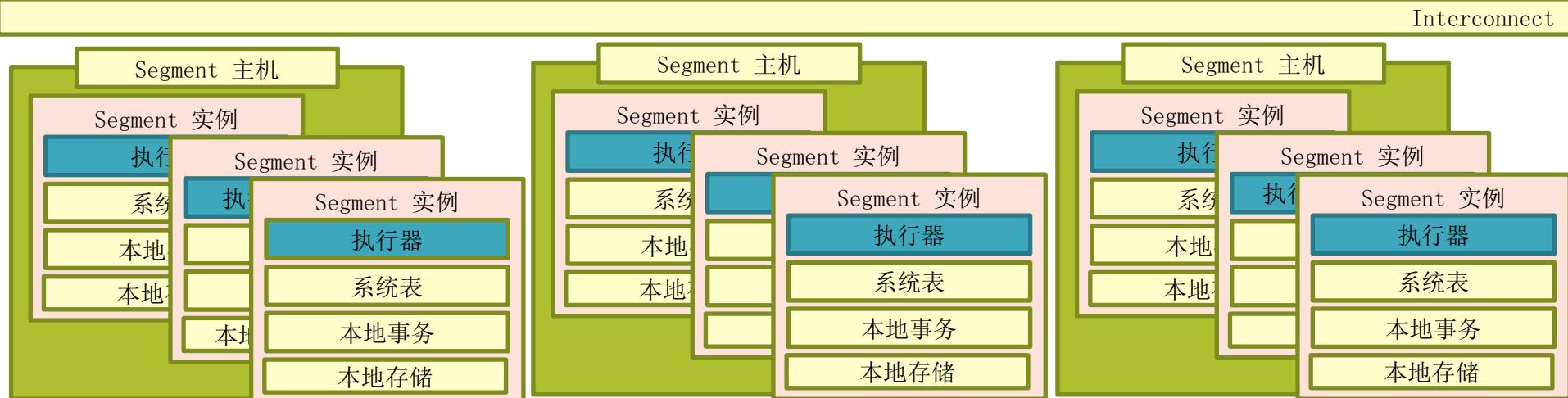
执行器

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

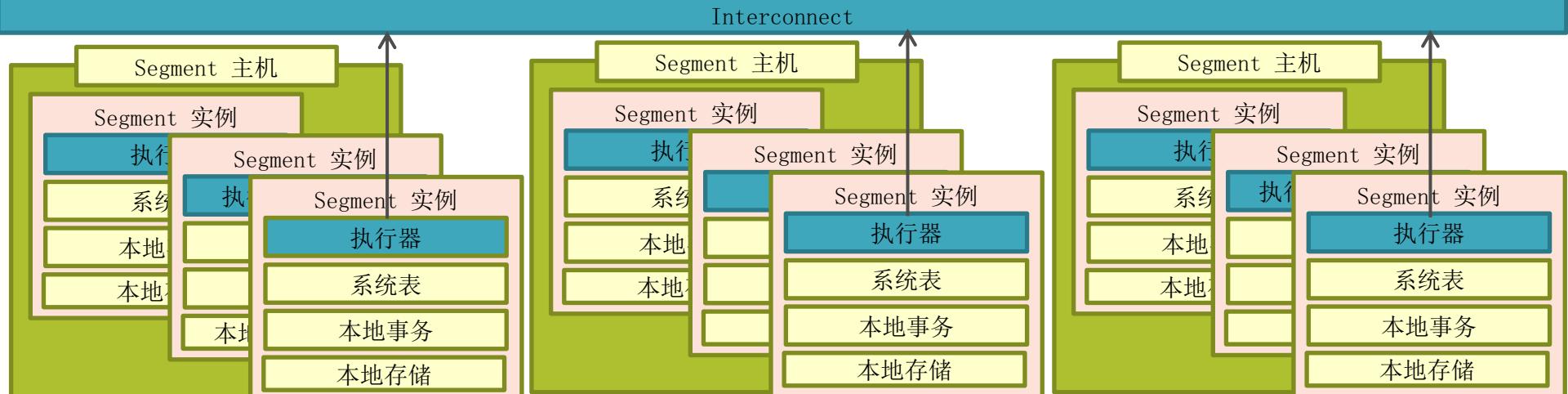


Interconnect

发送查询计划给各
个Segments

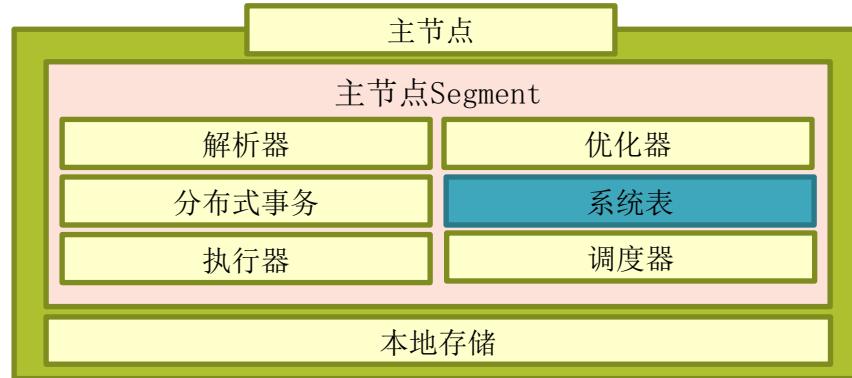


分配处理查询需要
的集群资源，收集
并返回结果给客户
端

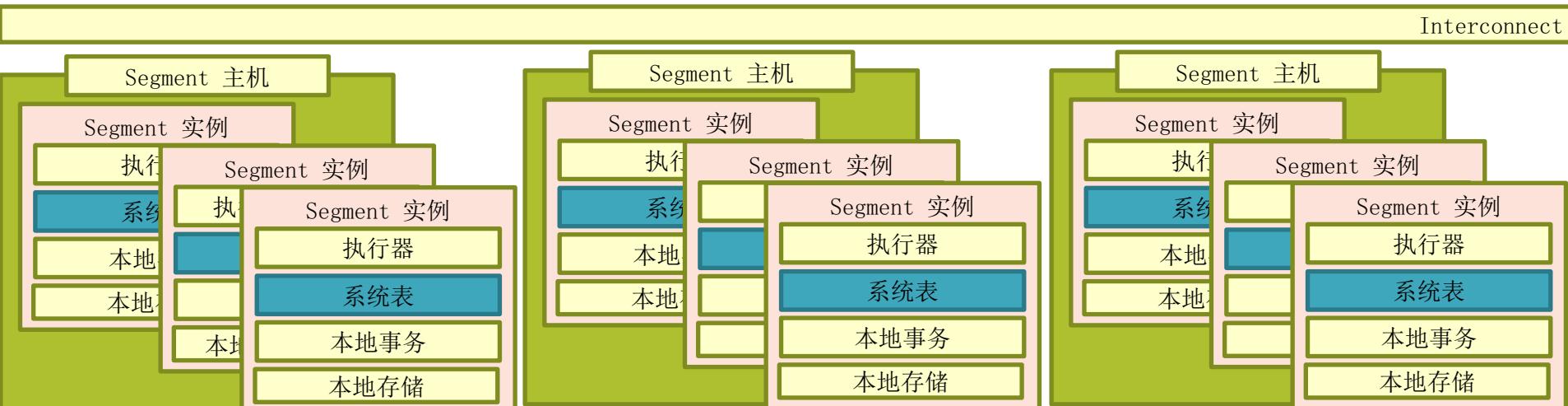


系统表

存储和管理数据库、表、字段的元数据

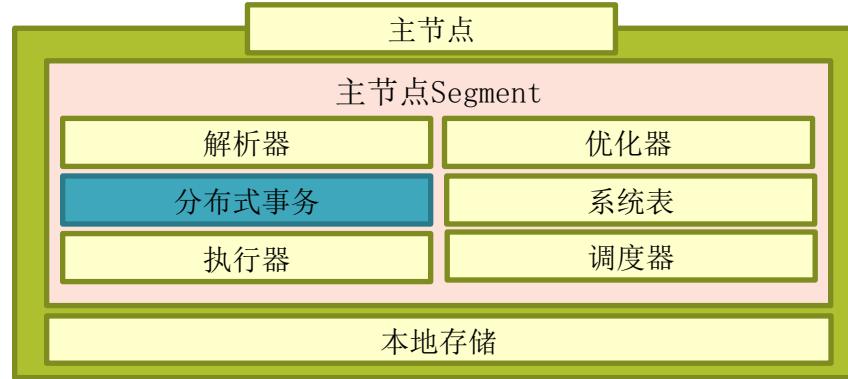


每个节点保存一个
拷贝



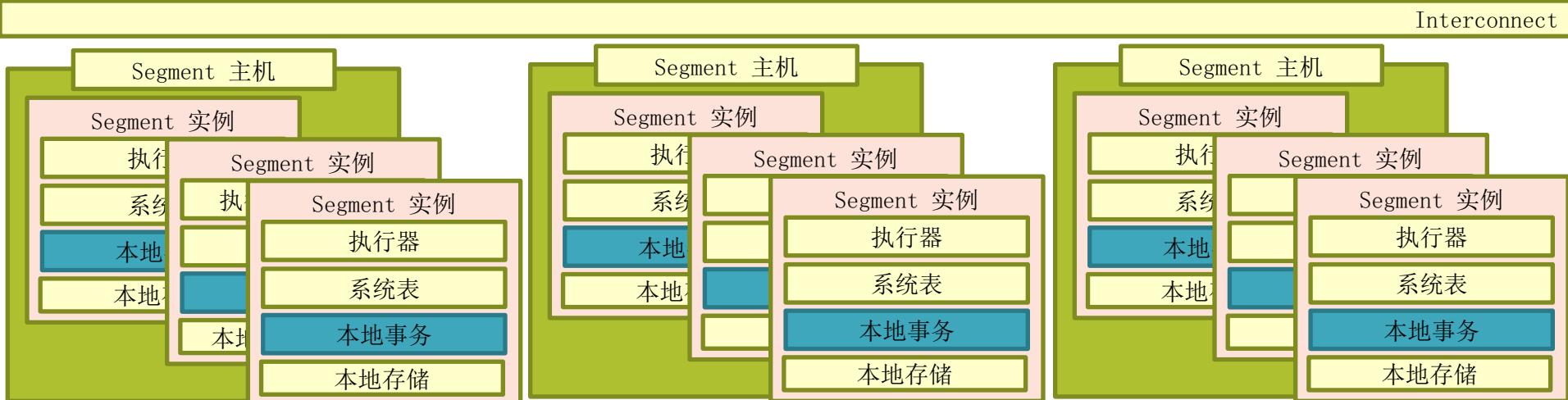
分布式事务

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



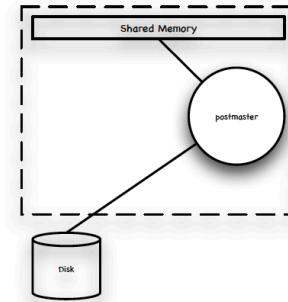
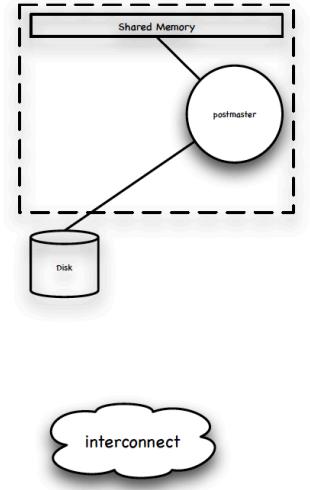
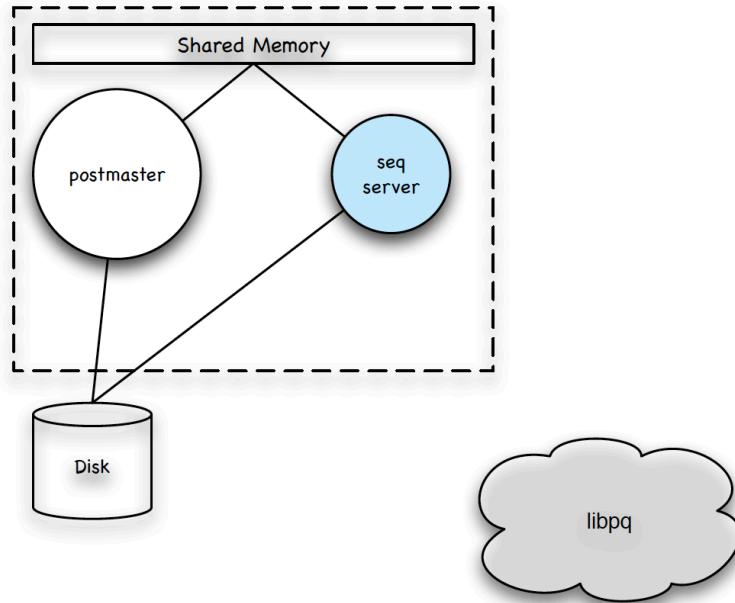
Segments 有自己的
事务日志，确定何
时提交或回滚自己
的事务

Interconnect

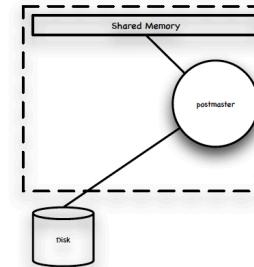
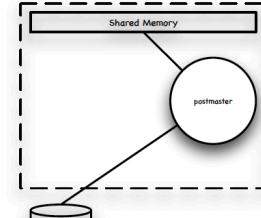
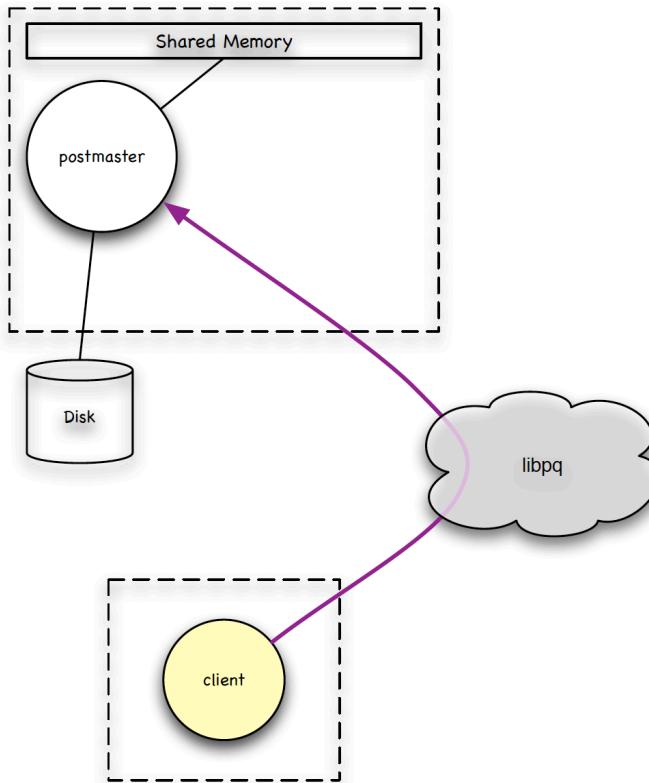


Greenplum 数据库SQL执行流程

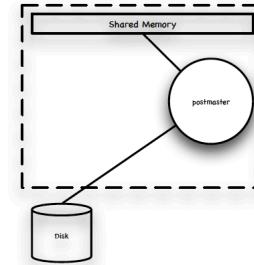
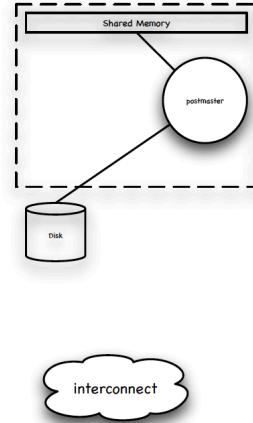
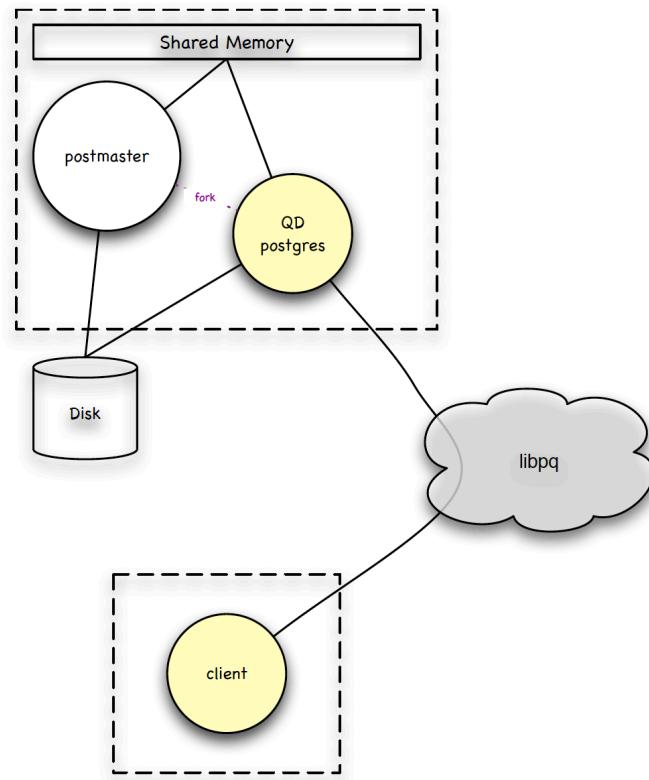
0. The system at rest.



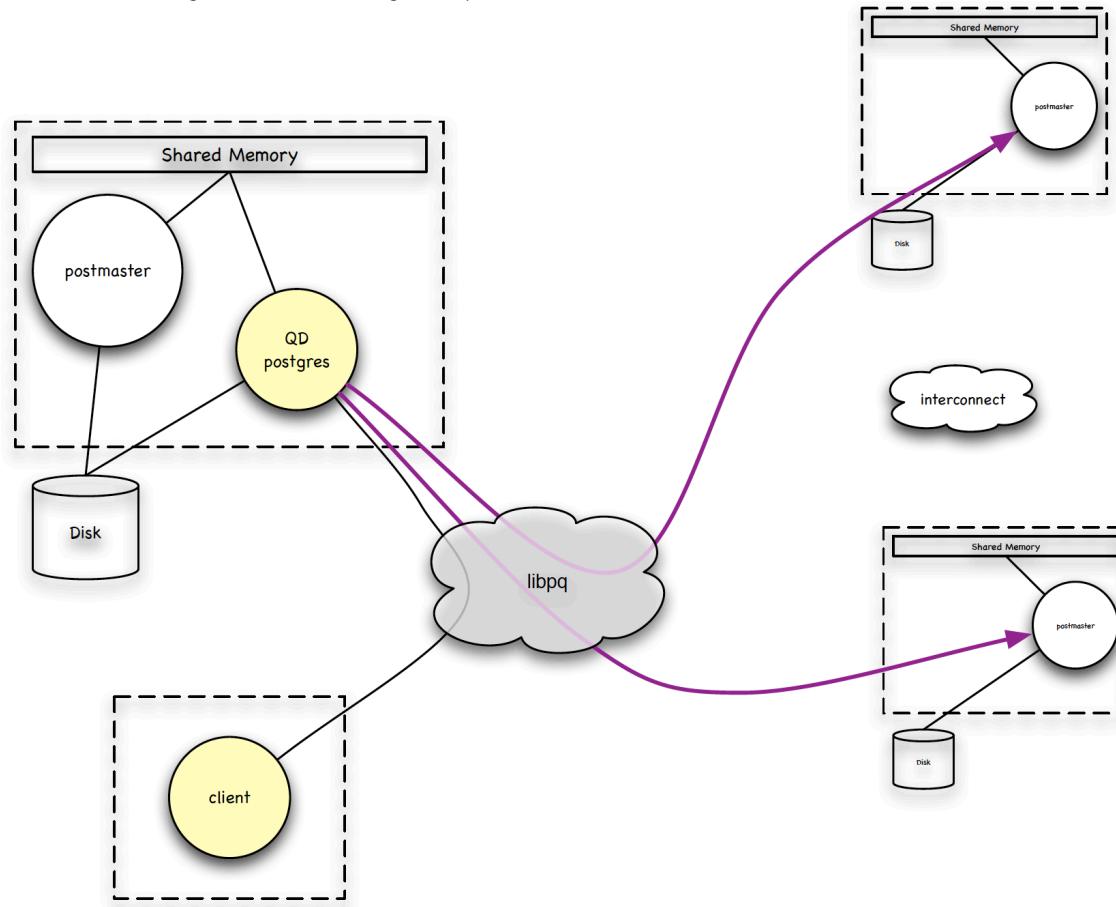
1. Client connects via the entry postmaster.



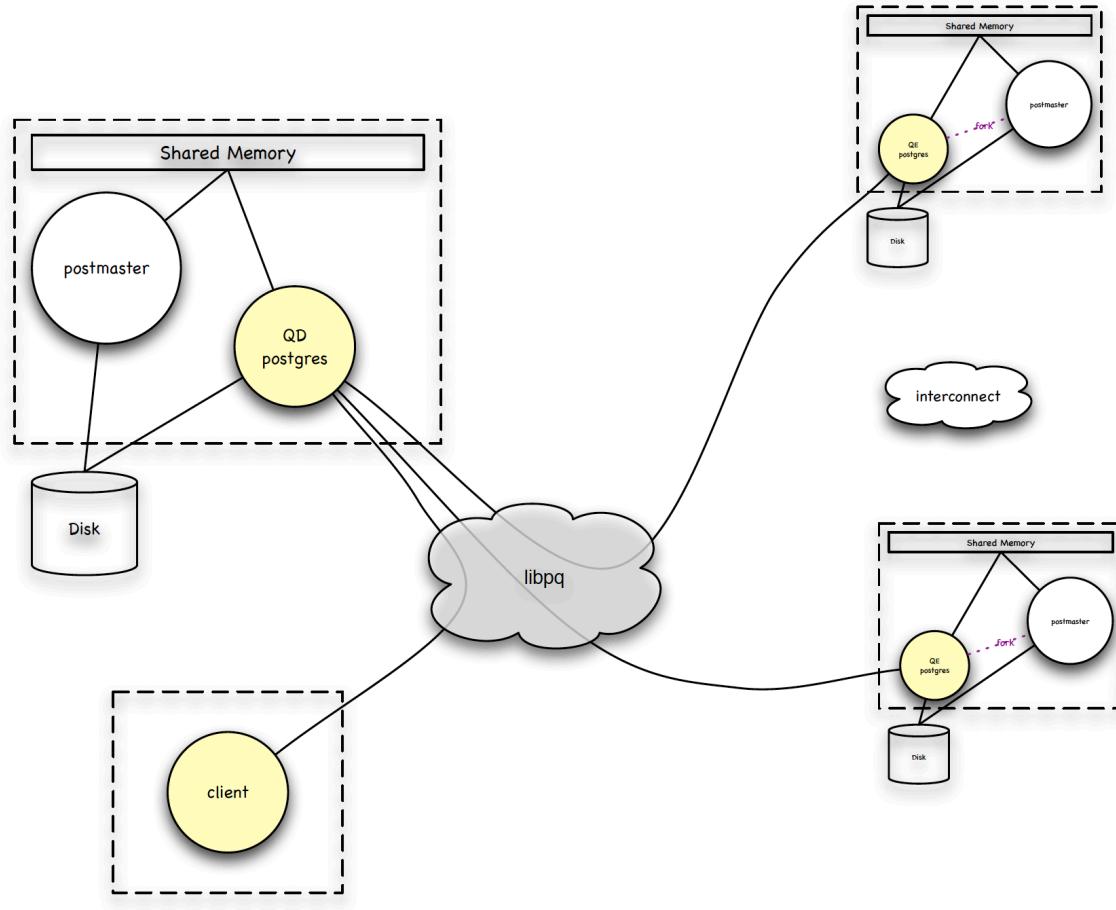
2. Entry postmaster forks a new backend -- the QD.



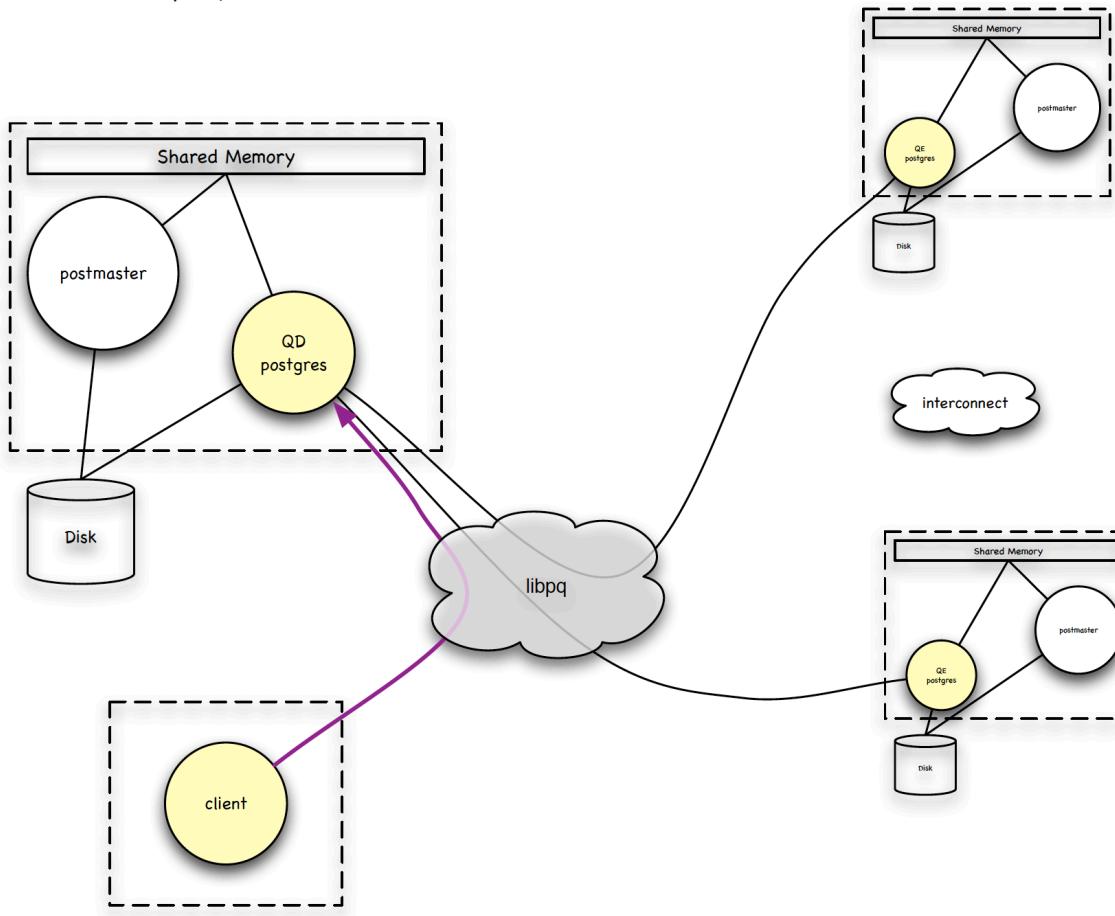
3. QD connects to segments via the segment postmasters.



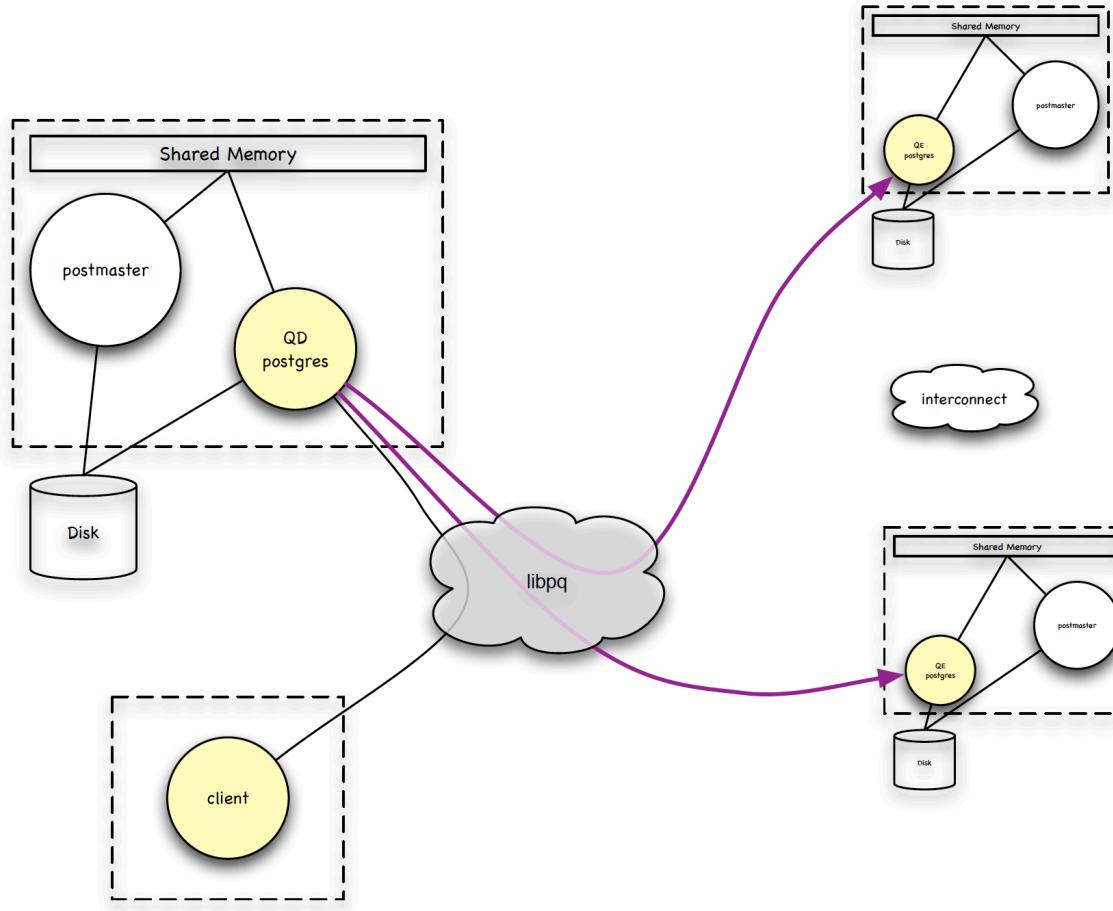
4. Segment postmasters fork initial gang of QEs.



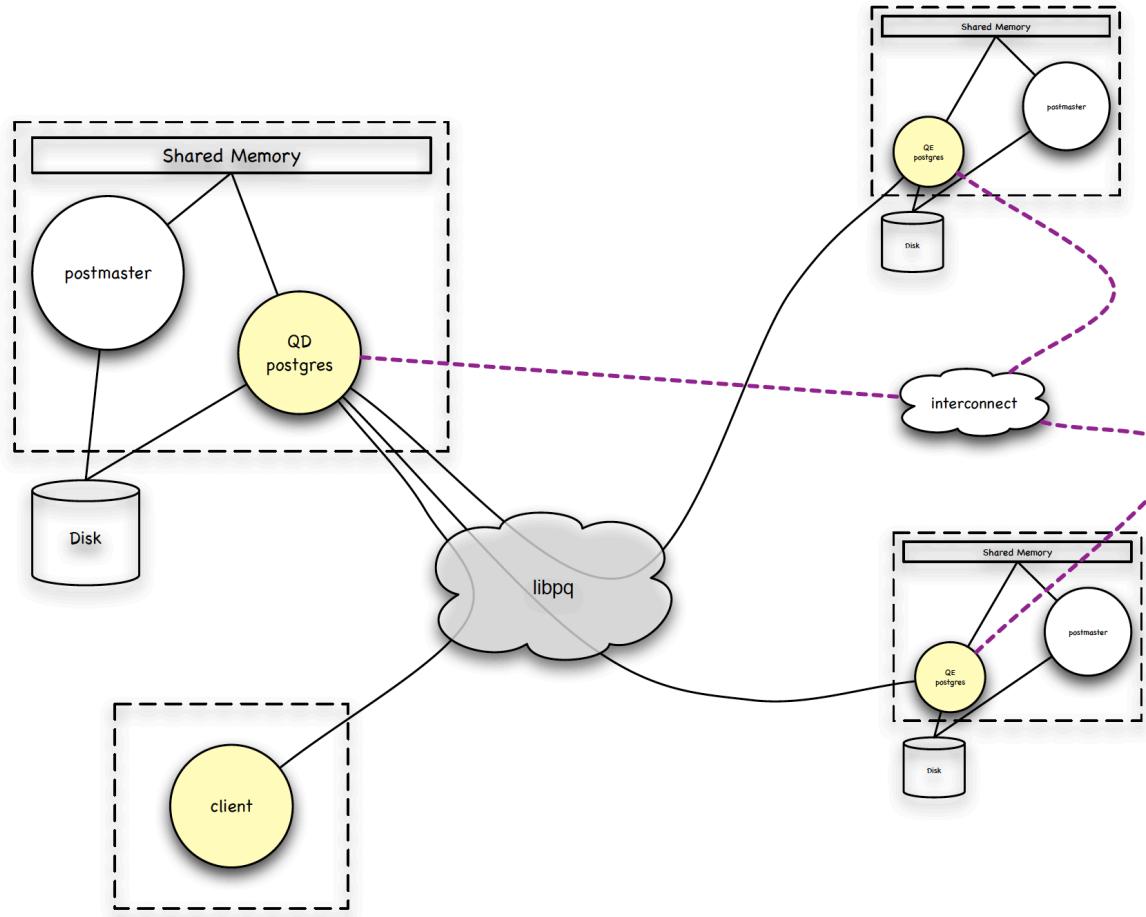
5. Client submits a query to the QD.



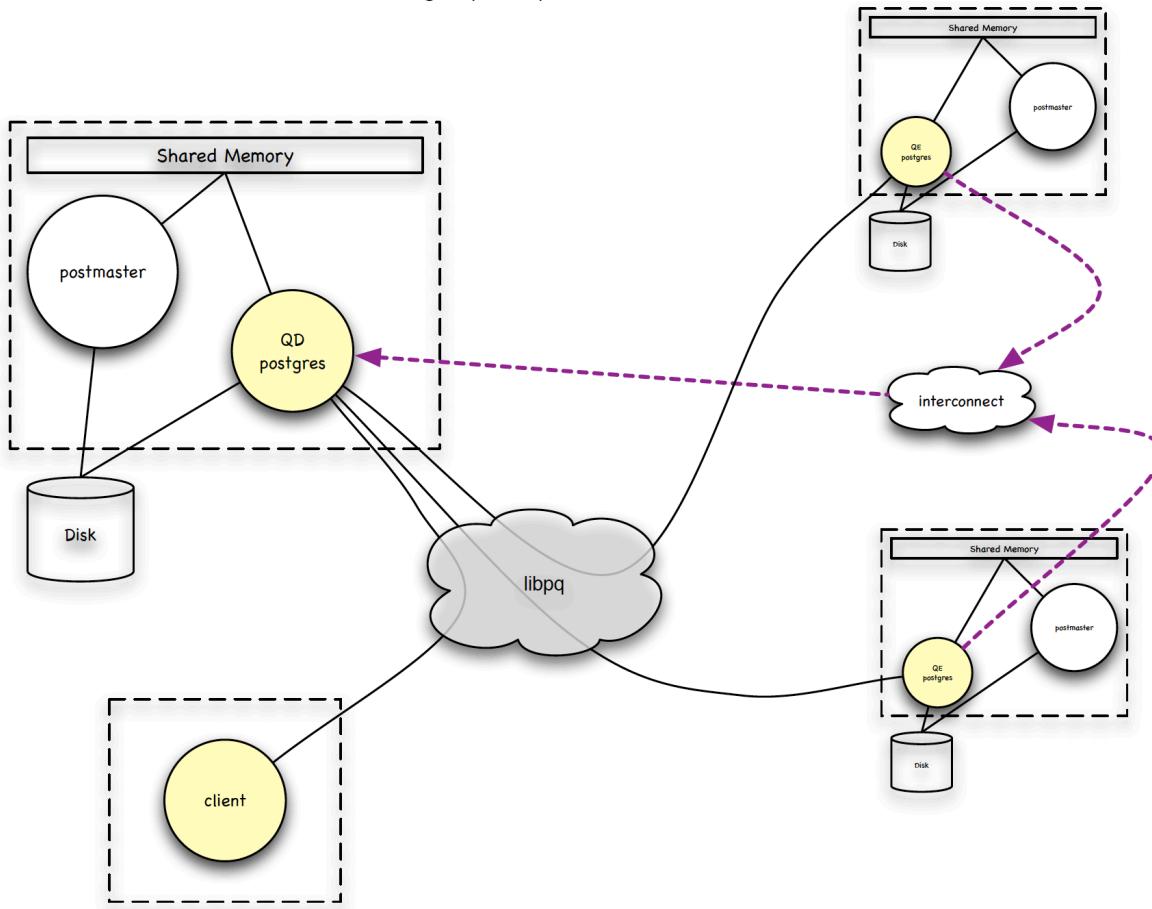
6. QD plans query and submits plans to QEs.



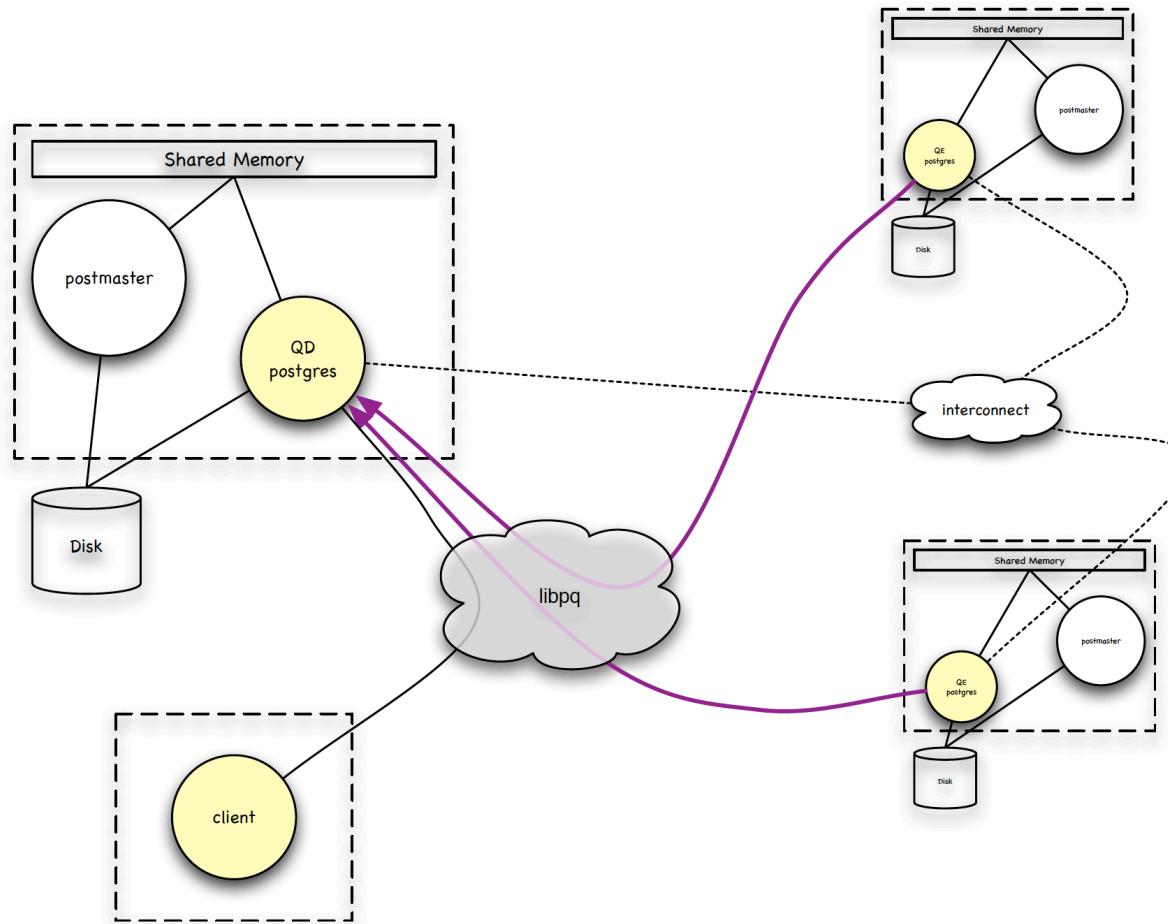
7. QD and QE's setup interconnect routes according to plan.



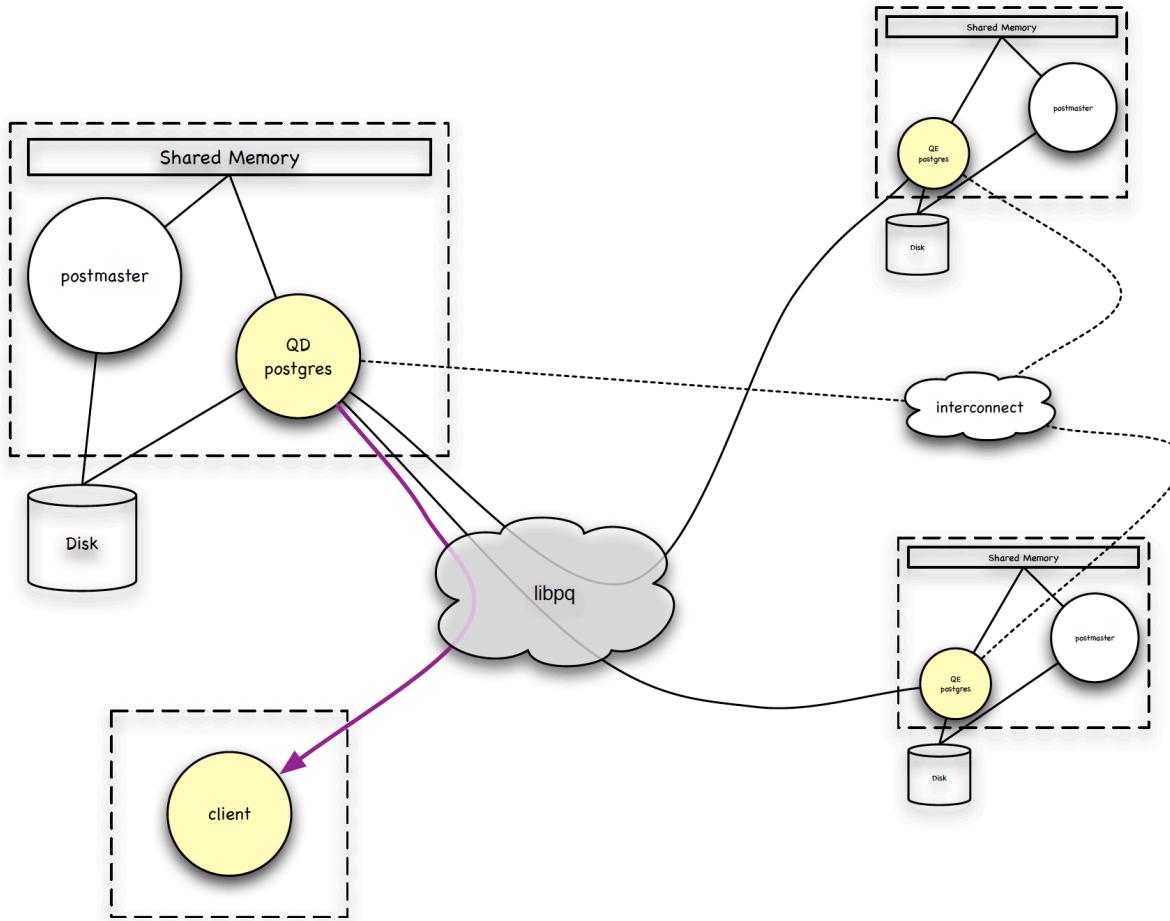
8. QD and QEs execute their slices sending tuples up the slice tree.



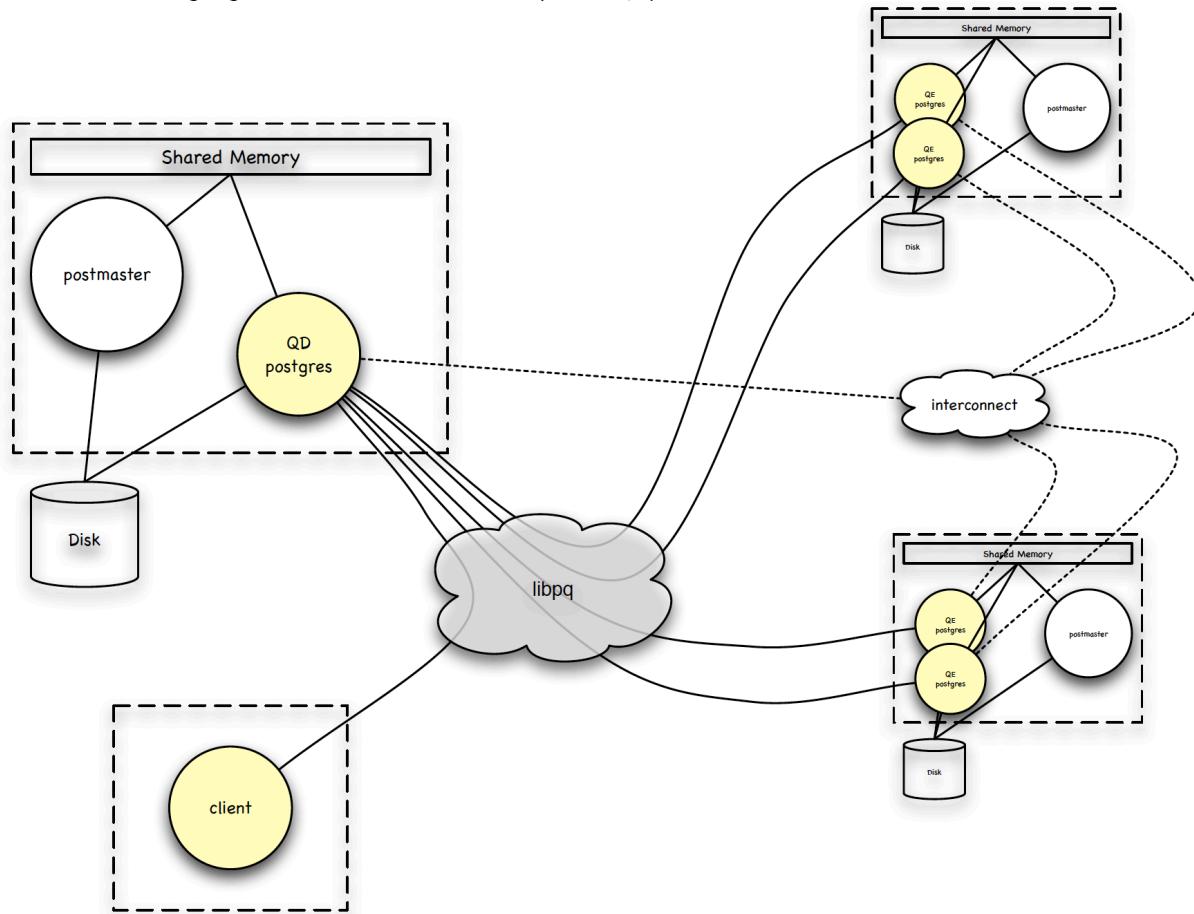
9. QEs return status to QD.



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

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://greenplum.cn>



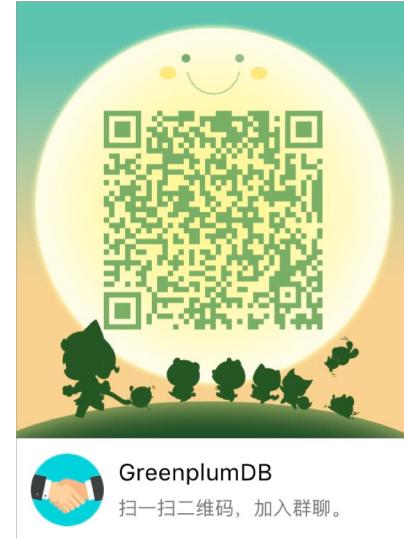
相关链接

- 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>



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
- QQ 群：



招聘

- 热招职位
 - Greenplum 数据库内核开发
 - Greenplum 文本检索产品开发
 - Greenplum GIS 产品开发
 - Greenplum 管理控制台前端和后端开发
- 联系方式:
 - pivotalrnd_china_jobs@pivotal.io
 - yyao@pivotal.io

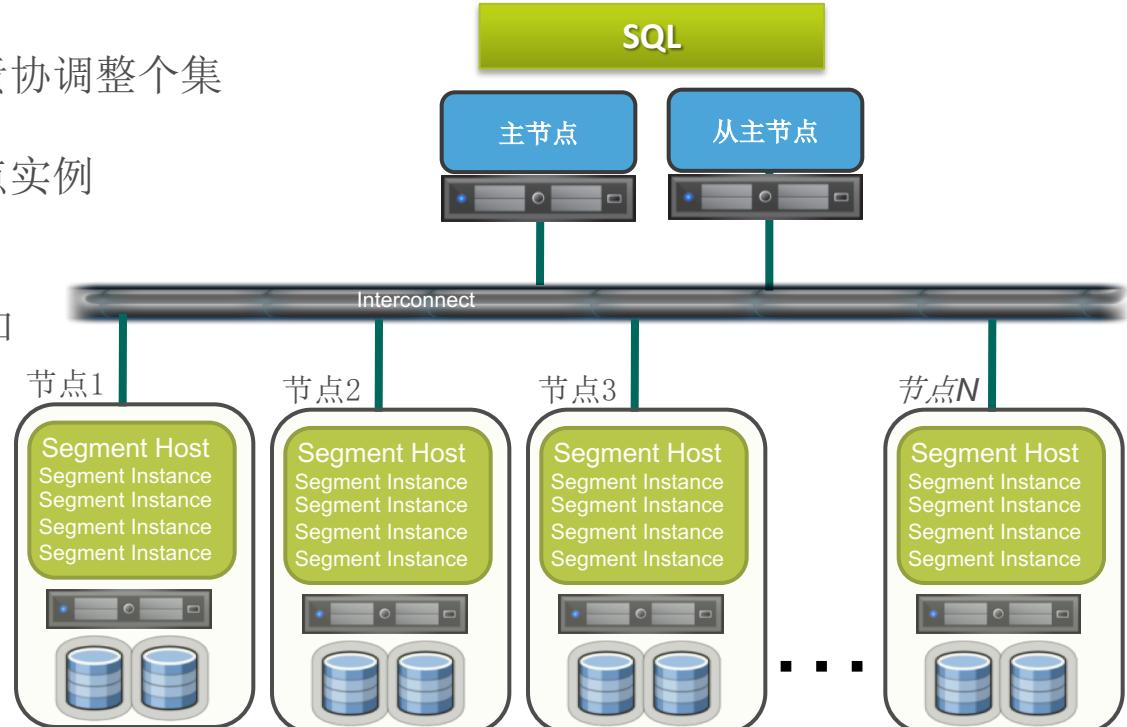


Appendix

高性能能源自并行化

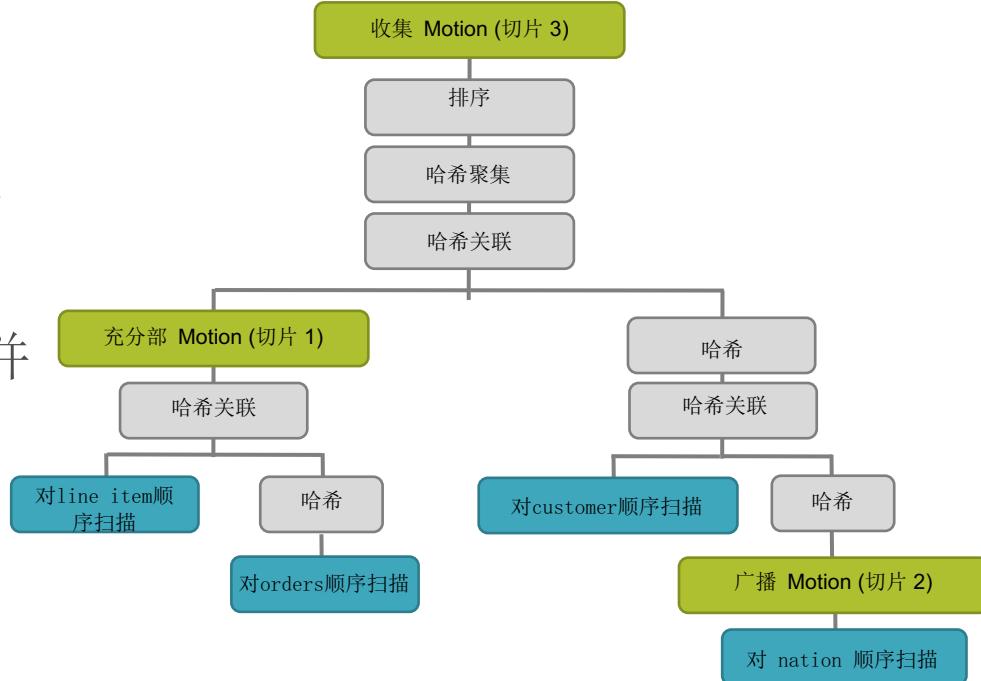
大规模并行处理、无共享架构

- 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）



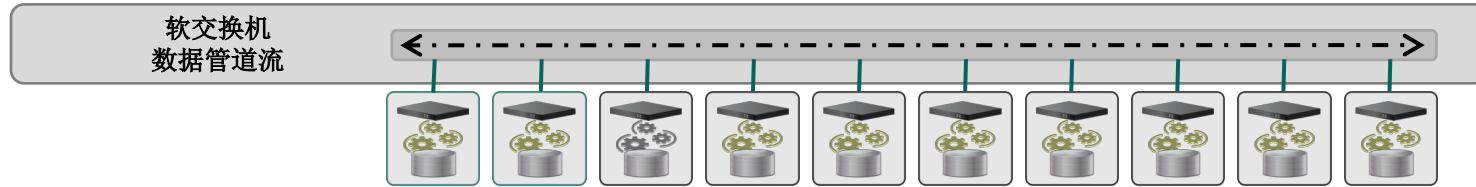
高性能能源自并行查询优化器

- 基于代价的优化器寻找最高效的查询执行计划
- 查询计划包括 scans, joins, sorts, aggregations 等多个切片
- 每个切片在各个Segment节点上并行执行
- Segment间通讯由Motion操作符处理



高性能源自动态数据管道流（Pipelining）

- 基于超级计算的“软交换机”负责：
 - 查询计划执行过程中在motion节点间高效处理数据流
 - 发送消息，移动数据，收集结果并协调系统中各个节点间的任务



GPDB 高可用性

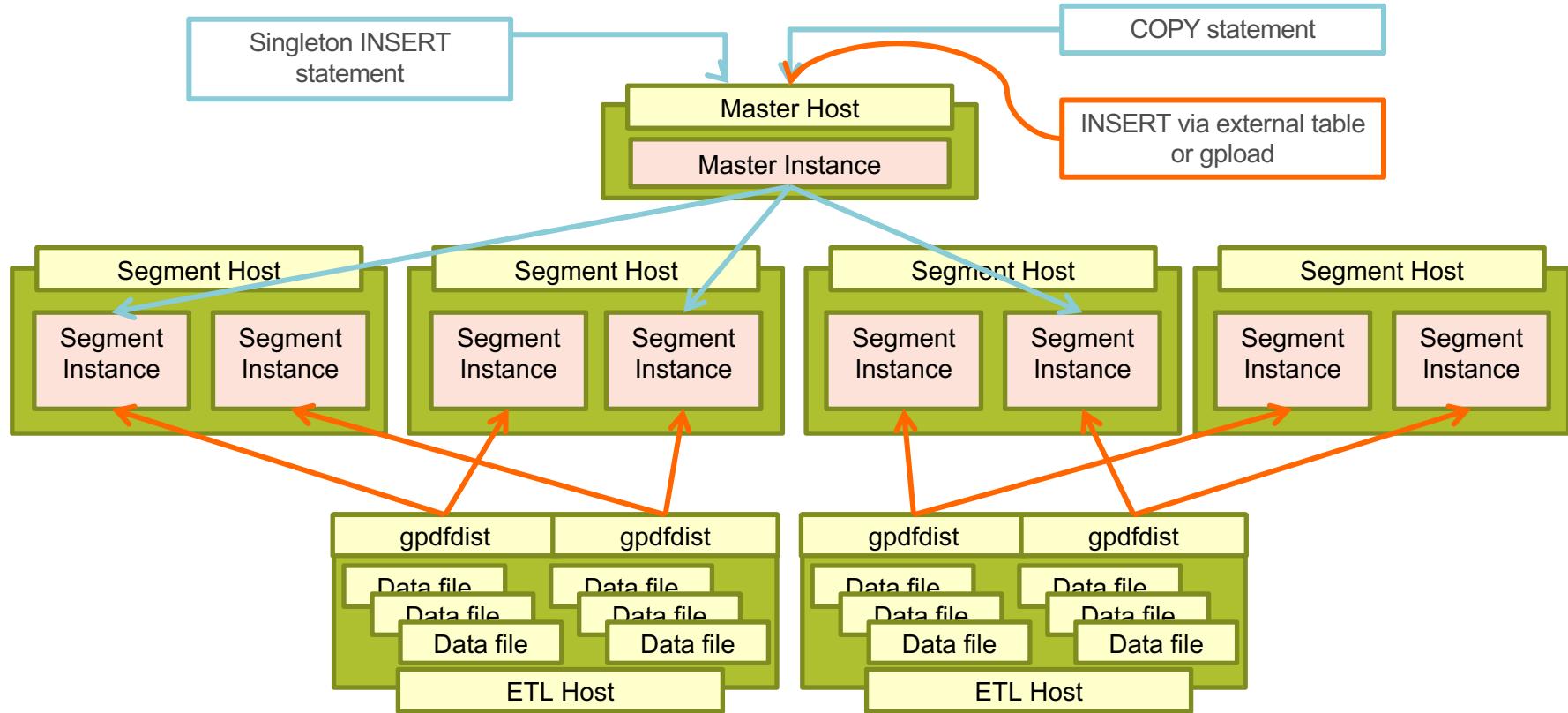
- 主节点（Master）高可用
 - Warm Standby
 - 主节点系统表副本
 - 避免单点故障
 - 主节点和从主节点间同步复制
 - 使用流复制
- 数据节点（Segment）高可用
 - 每个Segment都配备一个Mirror
 - 使用文件块级别的复制
 - 自动故障切换（failover）



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

Example Load Architectures



Critical Capabilities for Data Warehouse Database Management Systems

Published: 18 August 2014

Analyst(s): Roxane Edjlali, Mark A. Beyer

This Critical Capabilities research focuses primarily on existing customer implementations and their experiences with data warehouse DBMS products. Together with its companion Magic Quadrant, this report will help information leaders to select the right technology to meet their own use cases.