开源社
kaiyuanshe

2016中国开源年会

China Open Source Conference 2016

Greenplum 在线扩展

马涛

2016年10月

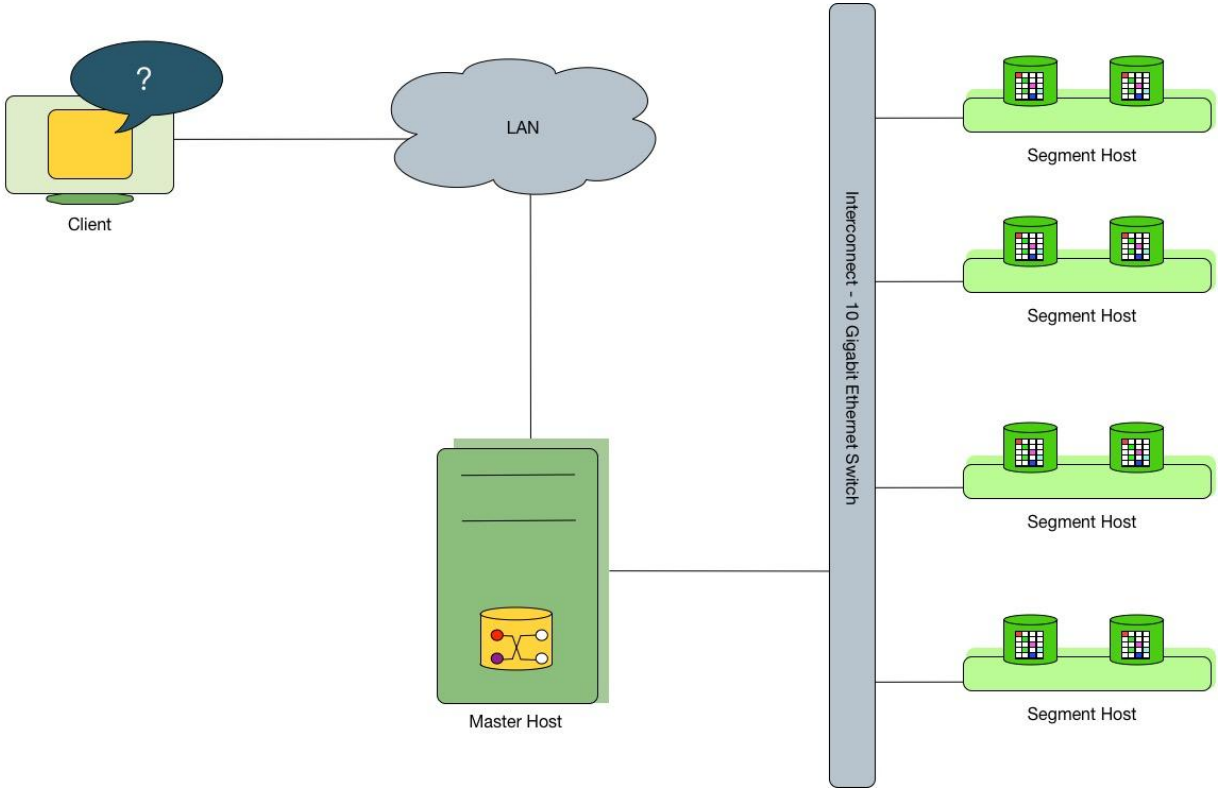
Agenda

- Greenplum Database 架构
- GP 在线扩展难点
- GP 在线扩展过程
- GP 在线扩展优缺点
- 对 GP 在线扩展的改进

Greenplum Database架构

- 数据分布
  - Hash 取模
  - Master 保存全局元信息
  - Segment 保存局部元信息和用户数据，Hash 取摸打散存储
- 查询执行方式
  - 数据与 Segment 绑定，根据查询访问的数据，进行最优化执行
  - 查询分成多个阶段，每个阶段之间通过流水线传输中间结果
  - 每个阶段内部根据 Segment 数量进行并行执行
- 可靠性
  - Master 通过传送 WAL 保证全局信息不丢失
  - Segment 通过物理复制将数据写到镜像节点
- 强事务支持
  - 支持 ACID
  - 通过两阶段提交保证

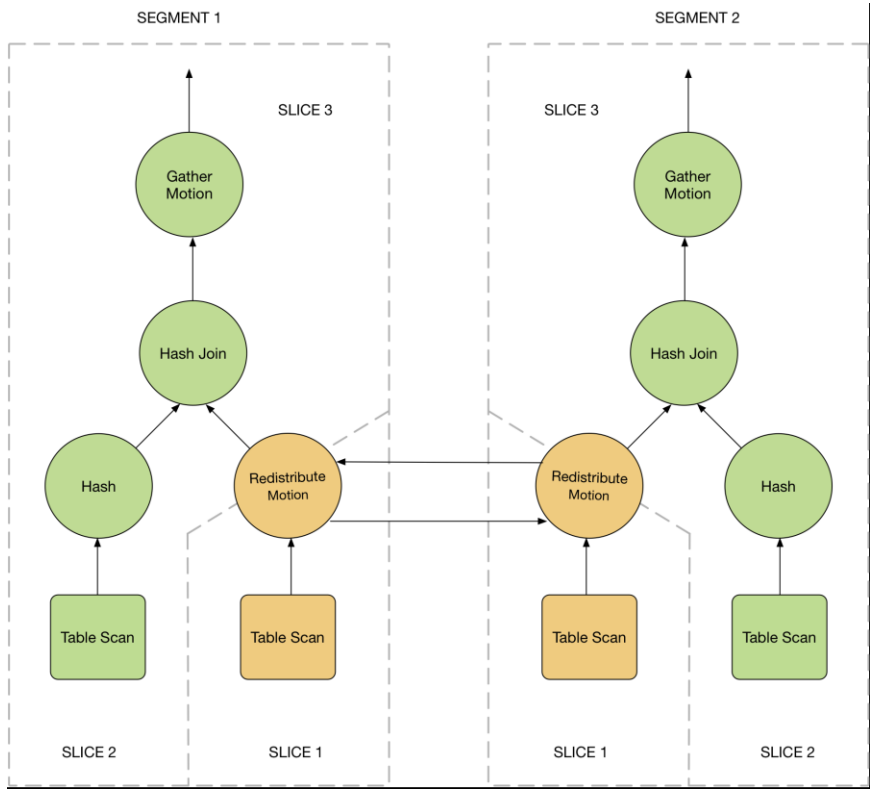
Greenplum 架构图



Greenplum 查询处理

- 一个简单的 SQL 查询：

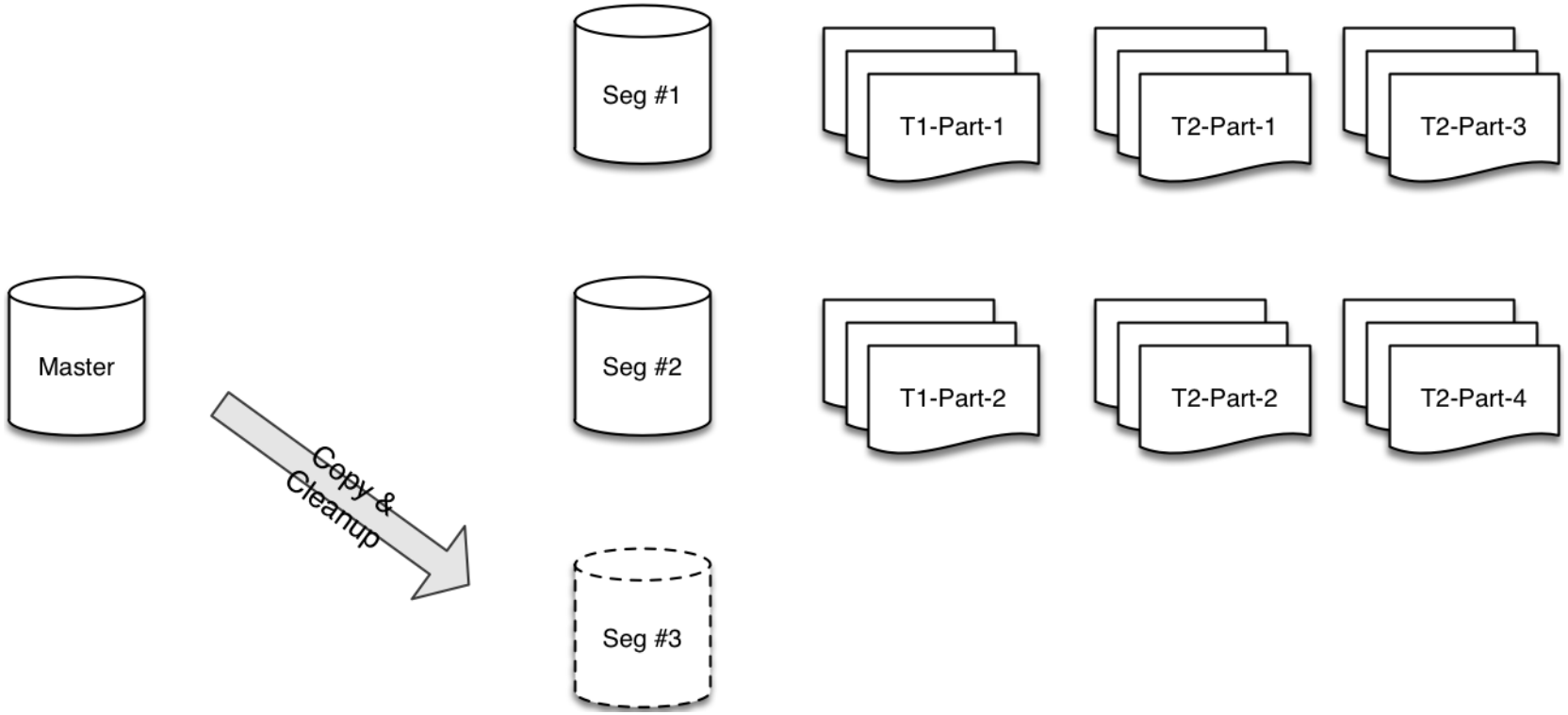
```
SELECT *
    FROM foo LEFT JOIN bar
    ON foo.id = bar.id;
```



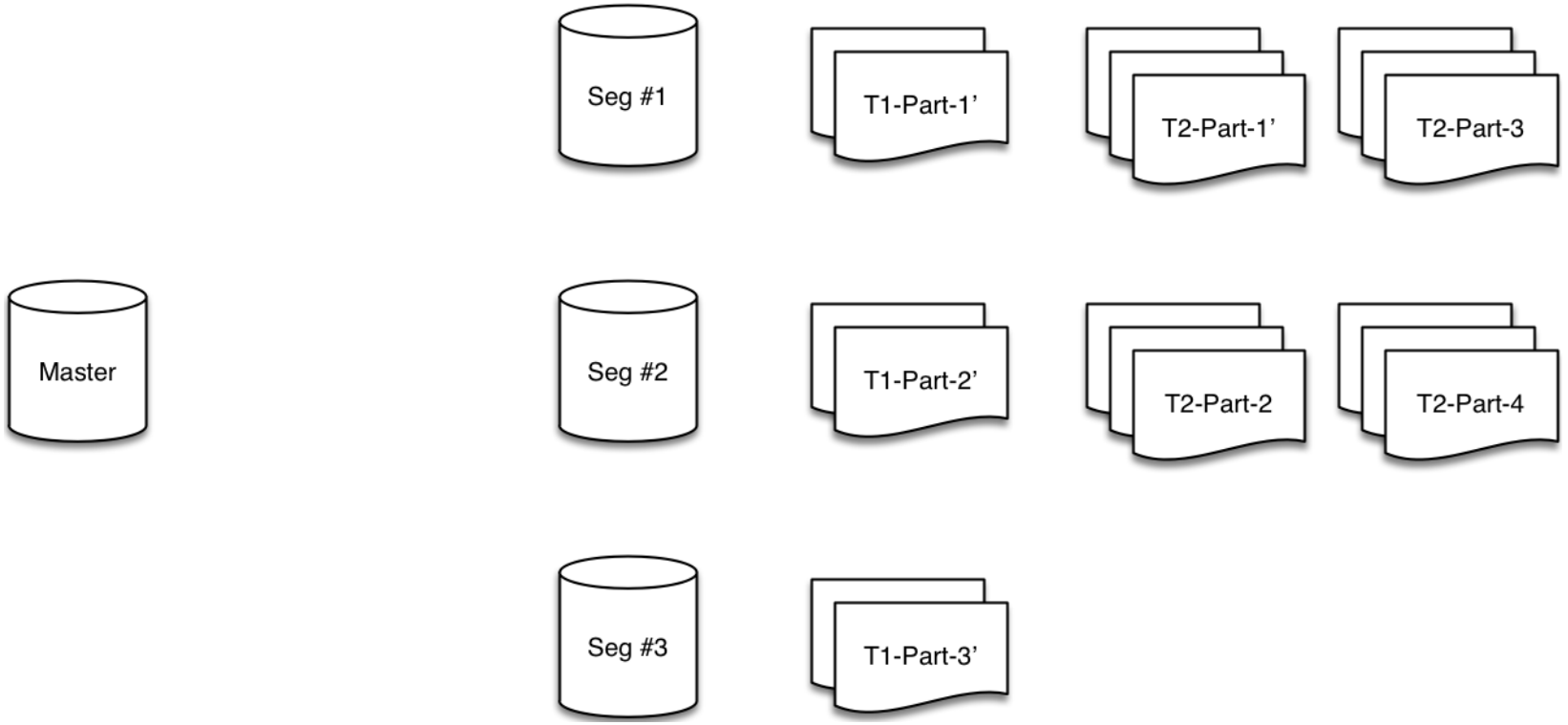
GP 在线扩展难点

- 数据分布
  - Hash 取模
  - Master 保存全局元信息
  - Segment 保存局部元信息和用户数据，Hash 取摸打散存储
- 查询执行方式
  - 数据与 Segment 绑定，根据查询访问的数据，进行最优化执行
  - 查询分成多个阶段，每个阶段之间通过流水线传输中间结果
  - 每个阶段内部根据 Segment 数量进行并行执行
- 可靠性
  - Master 通过传送 WAL 保证全局信息不丢失
  - Segment 通过物理复制将数据写到镜像节点
- 强事务支持
  - 支持 ACID
  - 通过两阶段提交保证

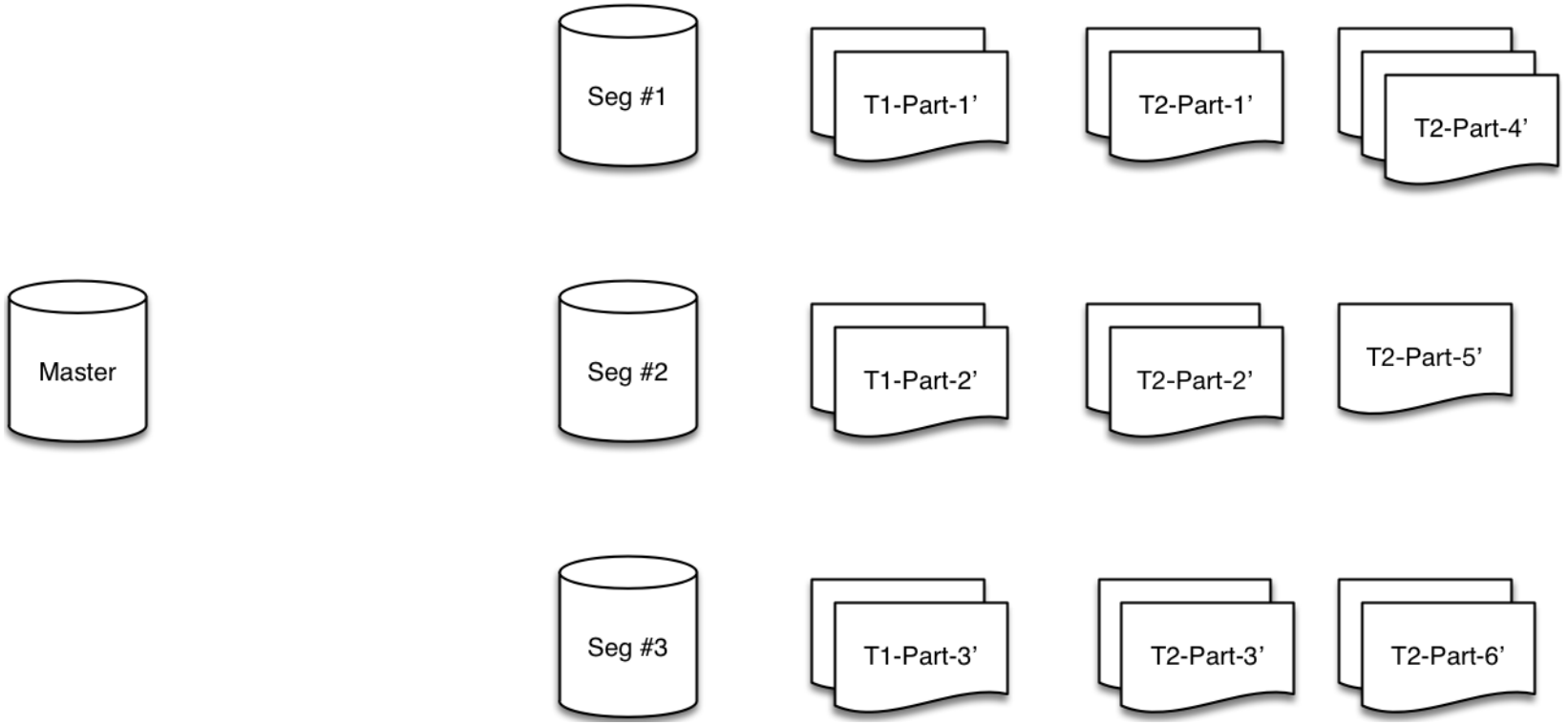
Greenplum 在线扩展过程1



Greenplum 在线扩展过程2



Greenplum 在线扩展过程3



GP 在线扩展的优缺点

- 优点
  - 计算存储耦合，查询性能非常好，查询时间开销稳定
  - 扩展过程保证强事务
  - 在线扩展过程可人工干预控制
- 缺点
  - 短时间服务中断
  - 数据重分布过程，查询性能下降
  - 所有数据都需要被读取和写入，I/O开销非常大

GP 在线扩展的改进

- 数据分布
  - 使用一致性 Hash 替代 Hash 取模（减少开销）
  - 移除 Segment 的局部元信息（避免停机）
- 查询执行方式
  - 优化器和执行器自适应集群规模和变化趋势

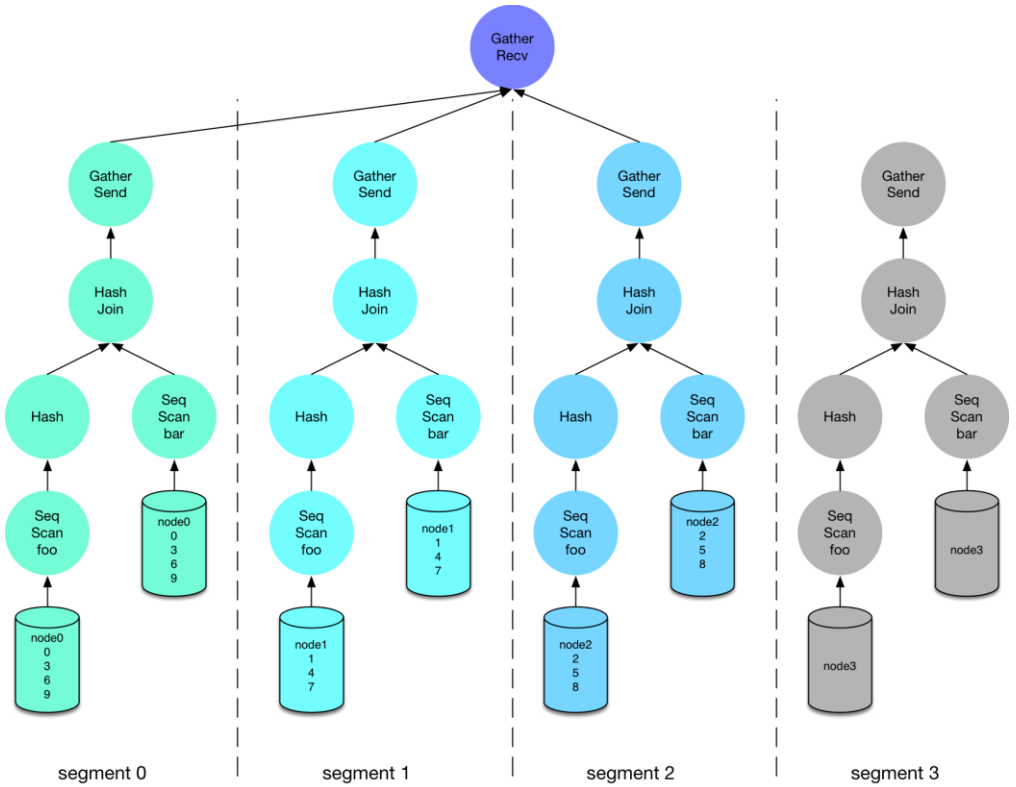
执行器改进(1)

- 一个简单的 SQL 查询：

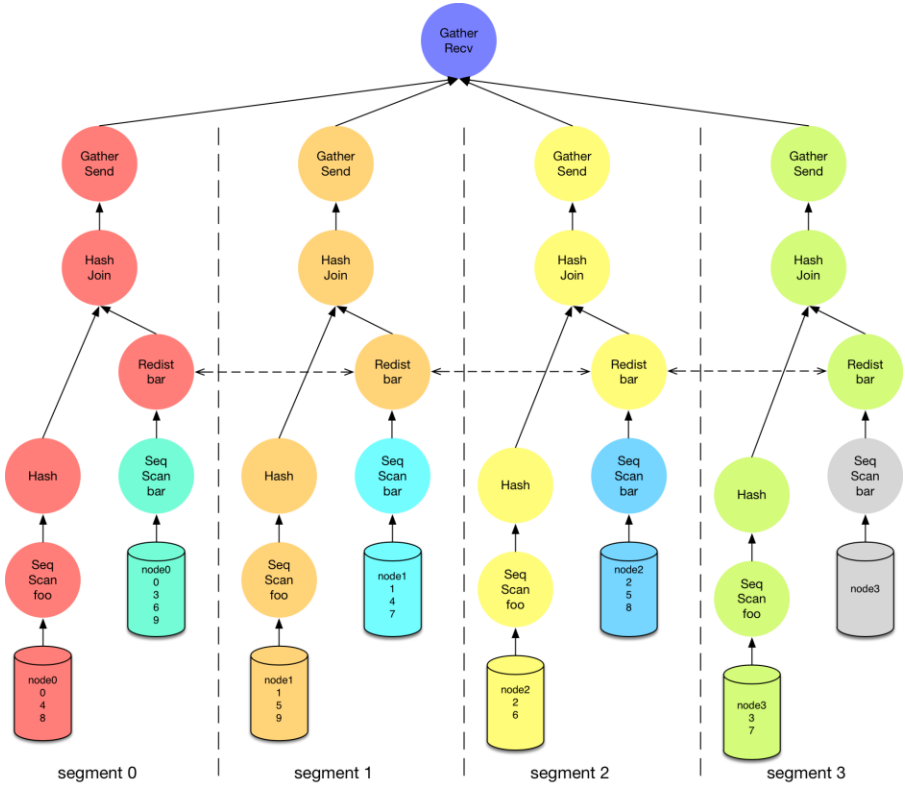
```
SELECT *
    FROM foo LEFT JOIN bar
    ON foo.id = bar.id;
```

表 foo/bar 都是根据 id 进行分片的

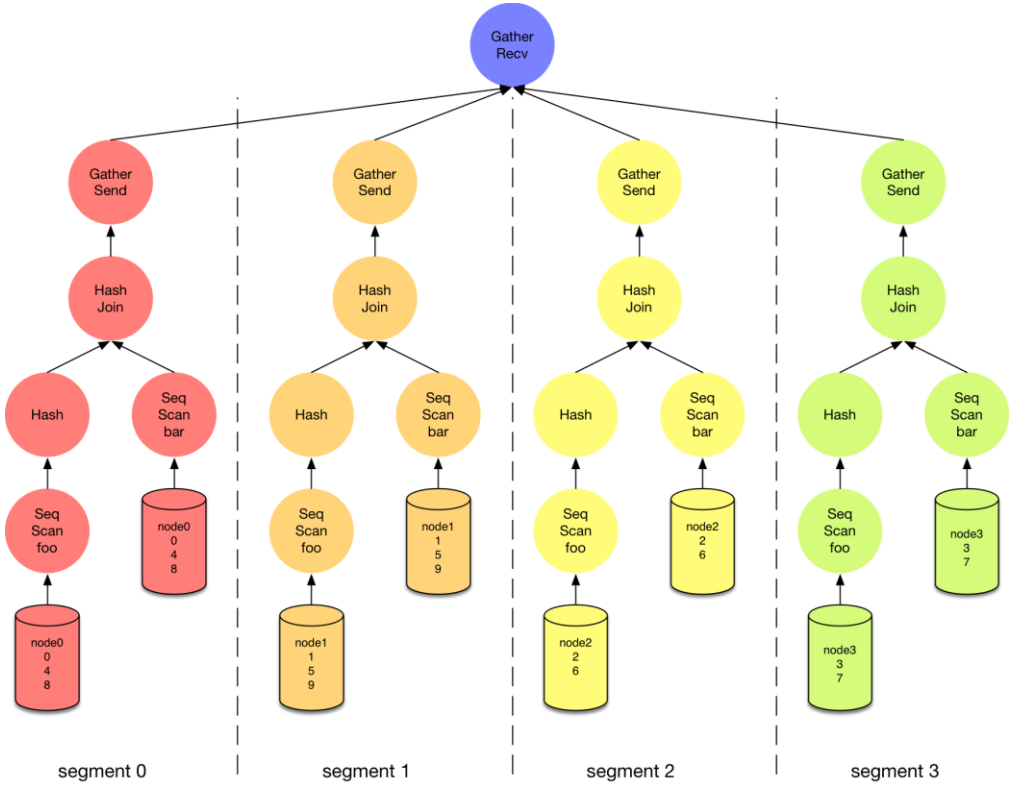
执行器改进(2)：两张表都没有重分布



执行器改进(3)：一张表都重分布完成



执行器改进(4)：两张表都重分布完成



在线扩展的重要性

- 私有部署
  - 耗时
  - 成本高
  - 性价比低
- 云计算
  - 计算资源无限
  - 花钱就能节省计算时间
  - 在线扩展允许集群资源动态增加

FAQ