

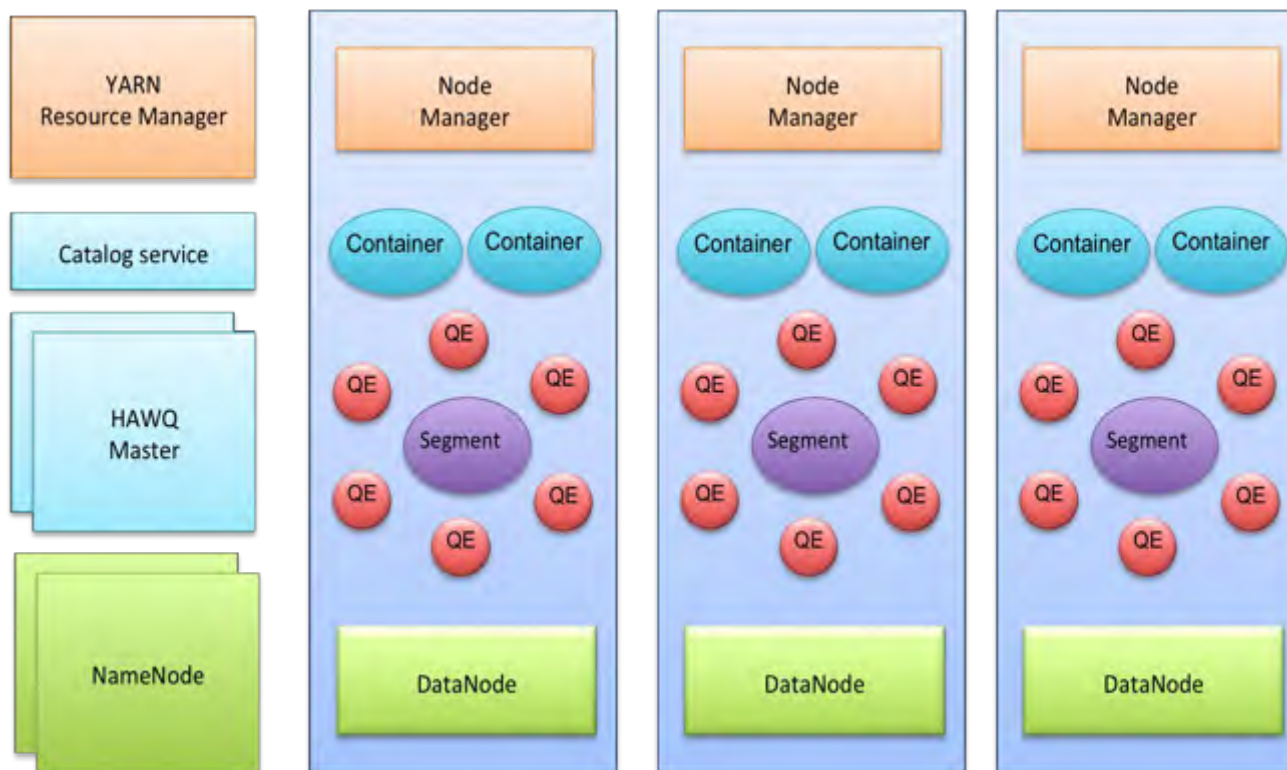
云时代大数据管理引擎HAWQ++

HAWQ简介及发展历程

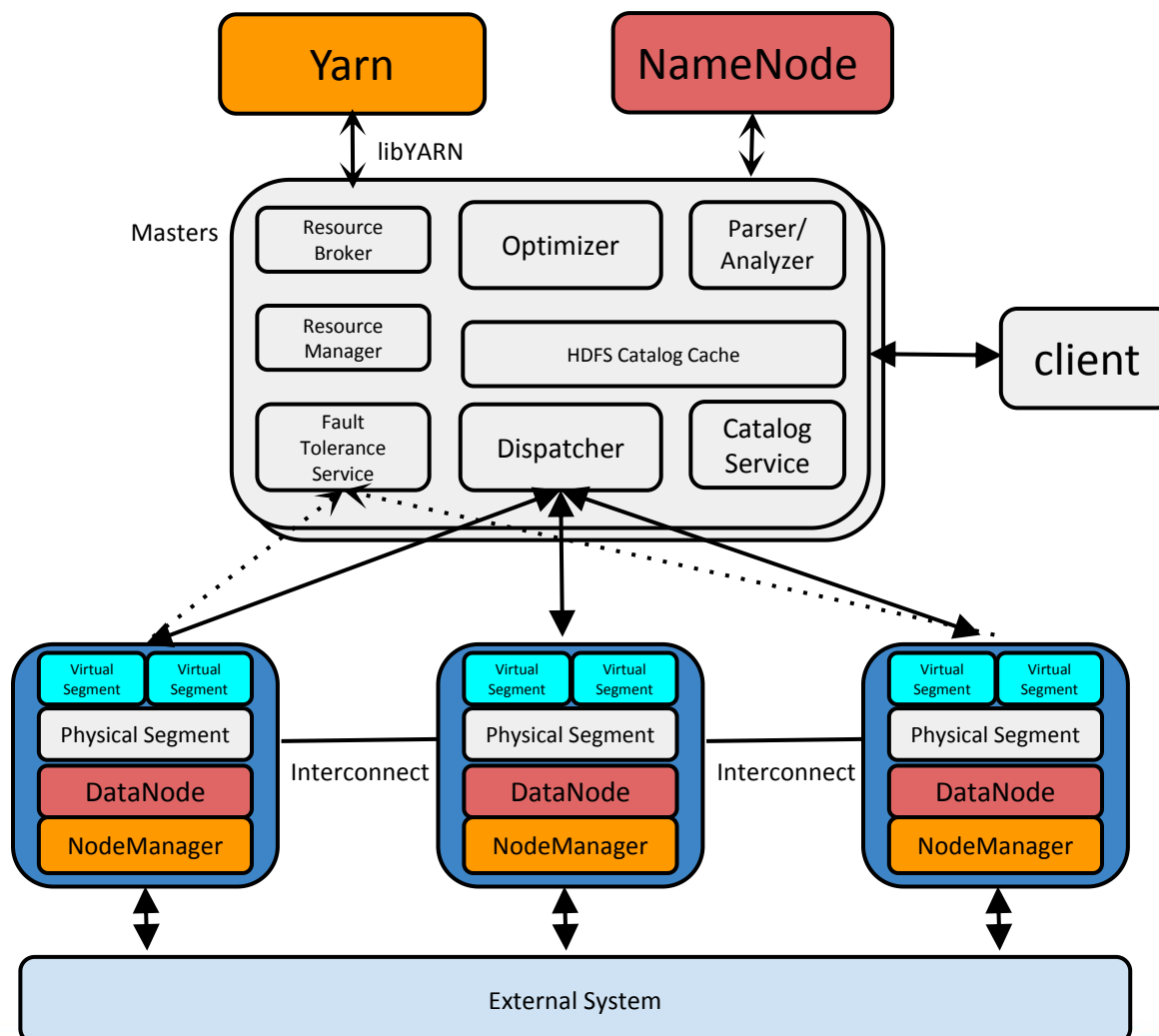
- 原生Hadoop并行SQL引擎
- GoH -> HAWQ Alpha -> HAWQ 1.0 -> HAWQ 1.x
- HAWQ 2.0 & Apache incubating
 - <http://hawq.incubator.apache.org>
- 偶数科技HAWQ++
 - <http://www.oushu.io>



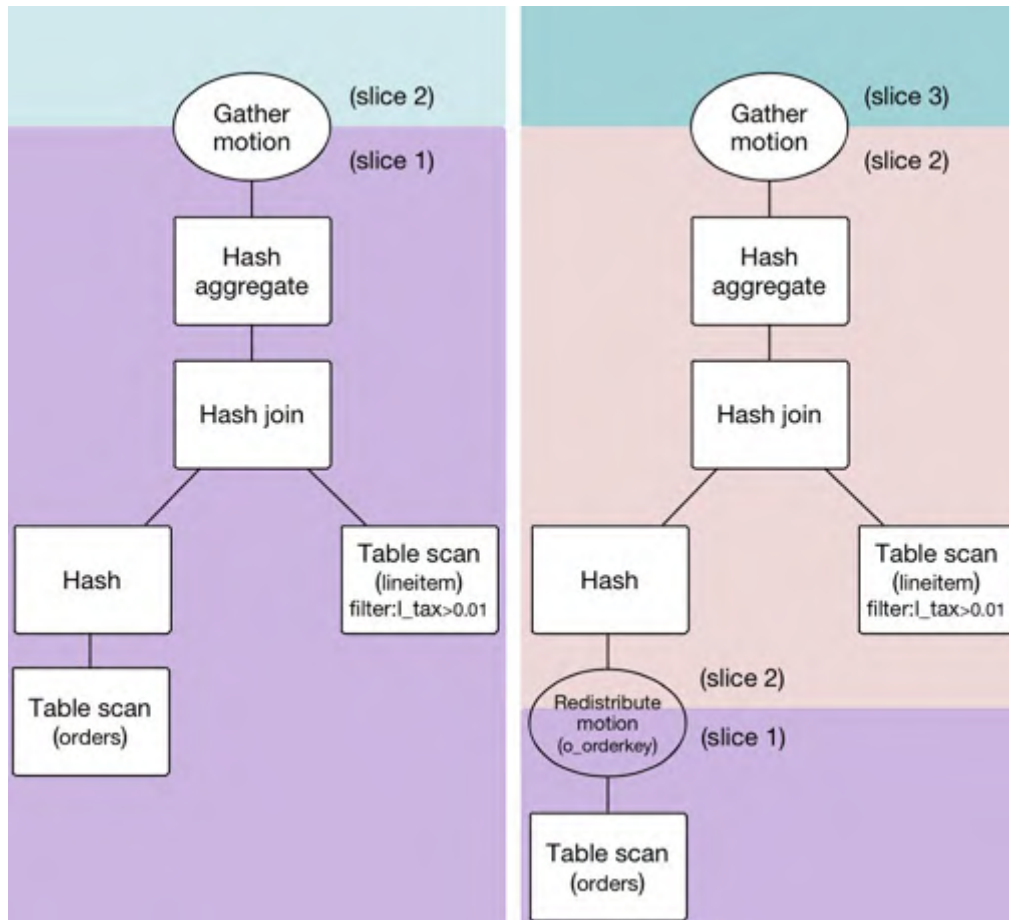
HAWQ组件



HAWQ体系架构



HAWQ优化器

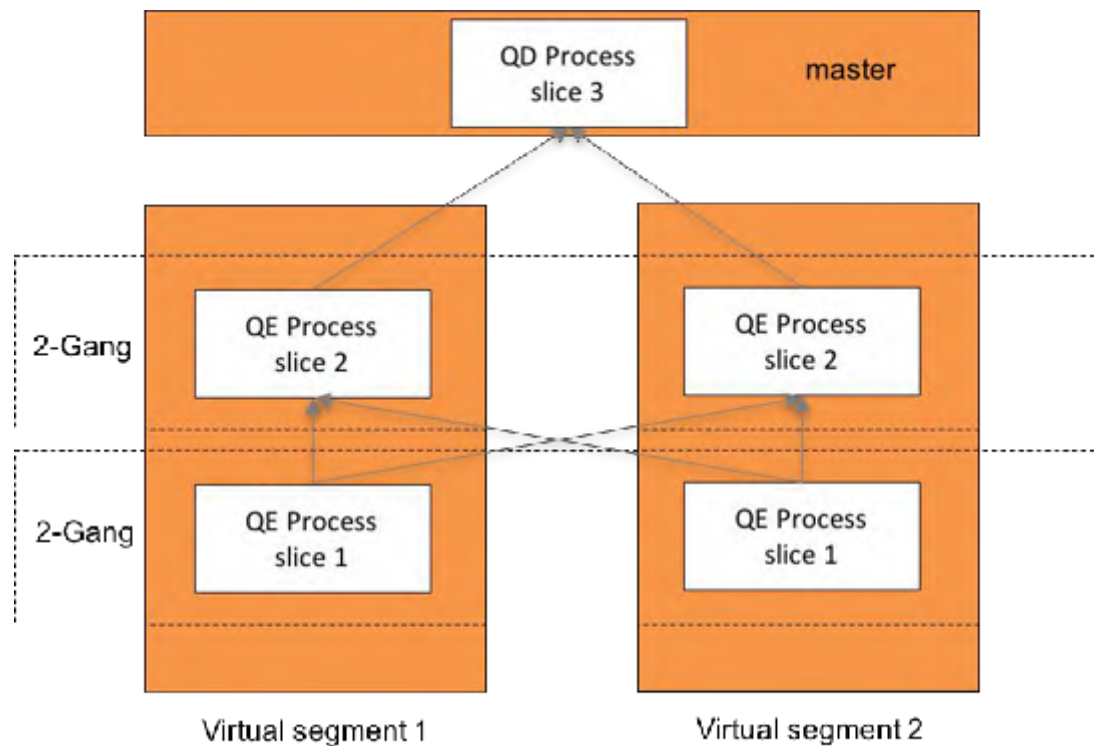


motion:

- **redistribute motion:** 按照 hash 值把数据发送到多个目标节点。
- **broadcast motion:** 把数据广播到所有目标节点。
- **gather motion:** 把数据从多个节点收集到一个节点。

```
SELECT l_orderkey, count(l_quantity)
FROM lineitem, orders
WHERE l_orderkey=o_orderkey AND l_tax>0.01
GROUP BY l_orderkey;
```

HAWQ查询处理流程



HAWQ资源管理

- 目标
 - 负责向YARN申请资源和回退资源
 - 负责为HAWQ用户，查询和操作符分配资源
- 三级资源管理
 - 全局资源管理
 - 内部资源管理
 - 操作符级别资源管理
- 多级资源队列
- CPU和memory的管理

HAWQ存储

- Row oriented: AO
 - Quicklz, zlib
- PAX like: Parquet
 - Snappy, gzip
- Other format: via PXF



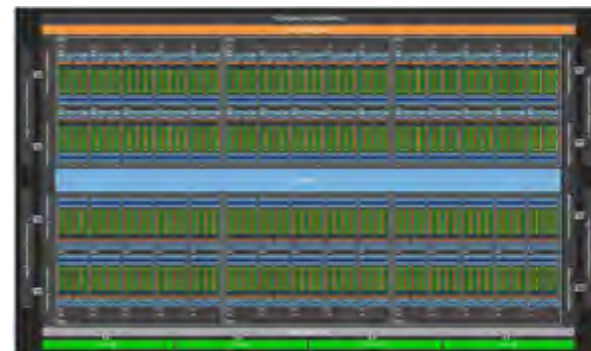
HAWQ++新特性

HAWQ++ 2.0.1& 2.1 (2017年Q1)

- 可插拔存储（文件系统，存储格式等）
 - 优化外部数据访问
- 容器云平台支持

HAWQ++ 3.0

- 高性能执行引擎 (已有原型系统)



HAWQ++安装部署

- 集群选择: 单节点 or 多节点
- 集群属性: HA, Kerberos
- 安装组件: Zookeeper, HDFS, HAWQ
- 安装方式: RPM手工安装 or Ambari部署集群

文档部分可以参考: <http://oushu.io/docs/hawq/>

Yum源: <http://yum.oushu.io/oushurepo/hawq++repo>

访问HAWQ++

- 交互式命令行工具psql访问HAWQ++
- libpq访问HAWQ++
- JDBC访问HAWQ++
- ODBC访问HAWQ++

HAWQ++和PostgreSQL兼容，可以使用PostgreSQL的驱动和编程接口访问HAWQ++

HAWQ++导入导出数据

- hdfs外部表导入导出数据

```
CREATE EXTERNAL TABLE ext_expenses ( name text, date date, amount float4, category text, desc1 text )  
LOCATION ('hdfs://nnhost:nnport/expense')  
FORMAT 'csv' (DELIMITER '|')  
LOG ERRORS INTO expense_errortable SEGMENT REJECT LIMIT 10 ROWS;
```

- gpfdist外部表导入导出数据
- COPY命令导入导出数据
- hawq load工具导入数据

HAWQ++结合MADlib的使用

- MADlib的安装
 - `gppkg -i <madlib_package>`
 - `gpstop -r`
 - `$GPHOME/madlib/bin/madpack -p hawq install`
- MADlib的使用：HAWQ++里默认的MADlib函数的schema是madlib

文档部分可以参考：<http://madlib.incubator.apache.org/>



谢谢！



陶征霖

偶数科技

www.oushu.io