




2016 杭州·云栖大会
THE COMPUTING CONFERENCE

云栖社区
yq.aliyun.com

MaxCompute SQL 2.0 全新的计算引擎

少杰
阿里云数据事业部 专家

2016
The Computing Conference

主办单位:  杭州

 Alibaba Group
阿里巴巴集团

战略合作伙伴: 



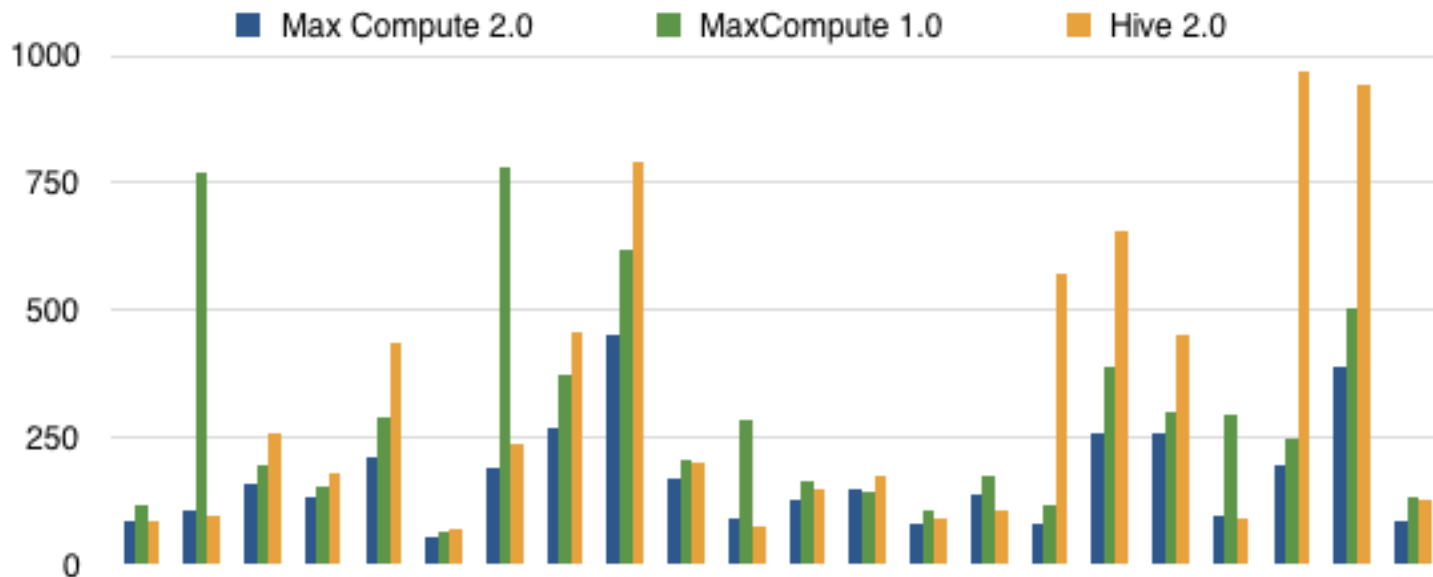
扫码观看大会视频

背景

- **MaxCompute SQL**
 - 分布式数据仓库
 - 批处理
 - 列存储
- **重大更新：2.0**
 - 全新的解析器
 - 全新的基于代价的优化器
 - 全新的运行时库
 - ...



TPC-H benchmark *



- VS Hive 2.0: +90%
- VS MaxCompute 1.0: +68%

* 1tb dataset on a 30-node-cluster



扫码观看大会视频

重大更新

全新的解析器	基于SQL的关系代数优化	基于代码生成的执行引擎
<ul style="list-style-type: none">• 基于ANTLR4重写的语法分析器• Playback实现更可靠的变更管理• 兼容Hive语法和语义*	<ul style="list-style-type: none">• 基于代价的优化器• 全新的基于代价的优化器 (Cost based optimizer)• 统计数据指导下的更精确的优化	<ul style="list-style-type: none">• 基于LLVM的高效代码生成• 向量化执行• 缓存友好的算法



基于代价的优化器

- 大多数现代DB实现了基于代价的优化器

Rule based	Cost based
Hive ($\leq 0.13^*$), SparkSql (Catalyst),	Oracle (≥ 7), SQL Server, MySQL, Postgresql, MaxCompute (≥ 2.0)

- 优势
 - 迭代的优化：搜索所有可能的优化路径
 - 动态规划：速度更快
 - 基于代价：更优的执行计划

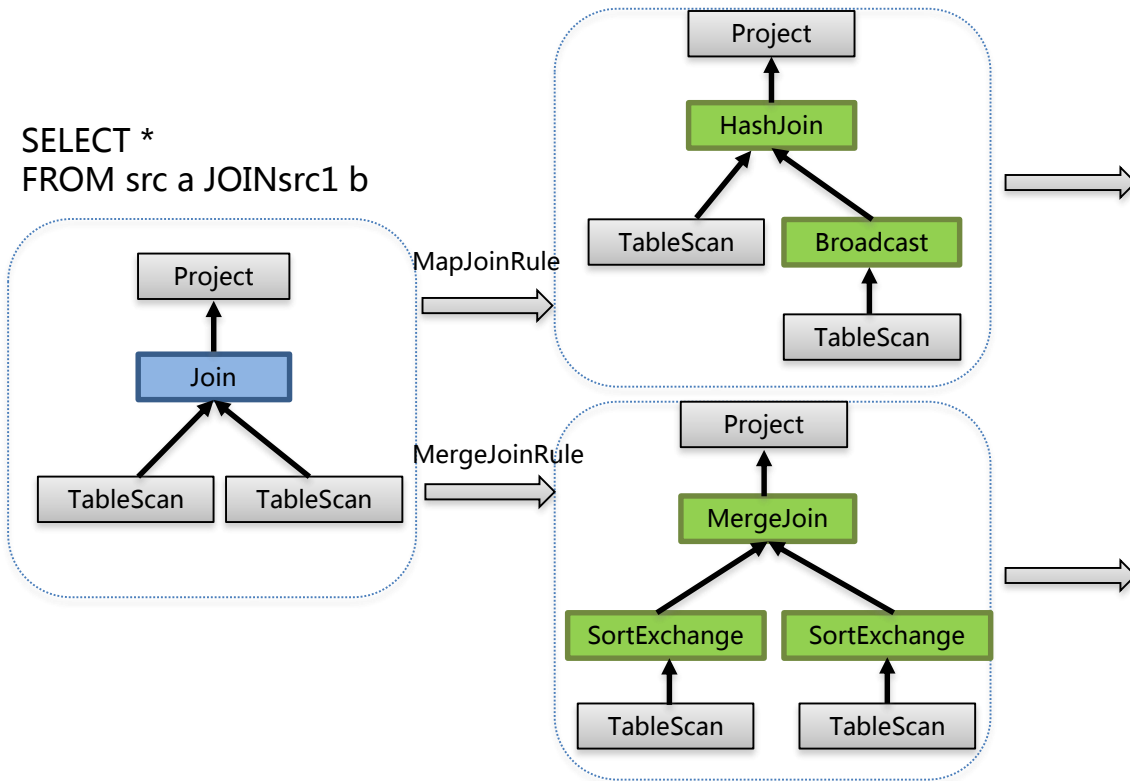


基于代价的优化器

- 实现方式

- 模式匹配
- 等价关系
- 动态规划的代价计算
- 最有计划搜索

SELECT *
FROM src a JOINsrc1 b



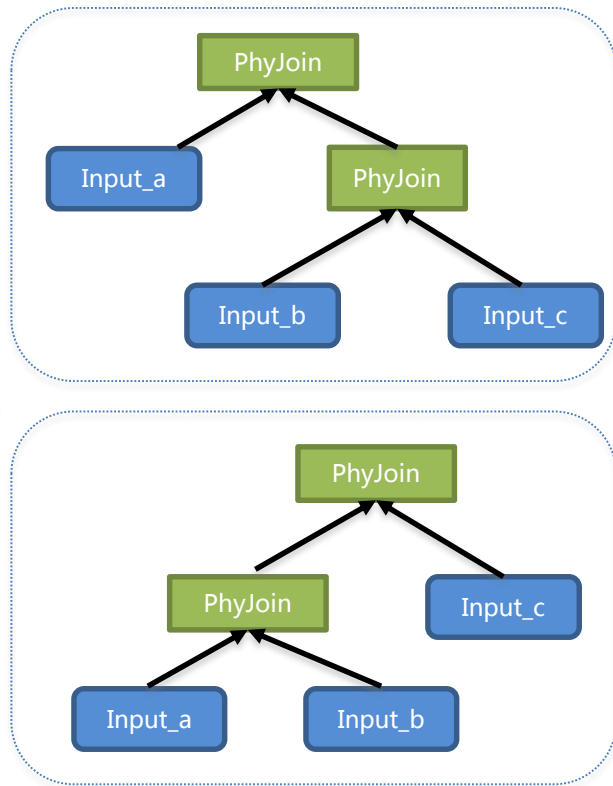
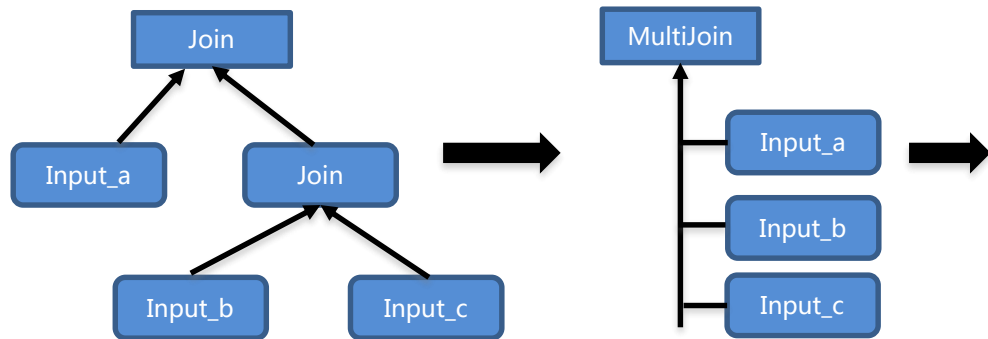
全新的优化规则

基础优化规则	裁剪	列裁剪/分区裁剪/子查询裁剪
	下推 / 合并	谓词下推
	去重	Project去重 / Exchange去重 / Sort去重
	折叠	常量折叠 / 谓词推导
探测优化规则	Join	BroadcastHashJoin / ShuffleHashJoin / MergeJoin SkewJoin
	Aggregate	HashAggregate / SortedAggregate / De-duplicate
	下推	GroupBy下推 / Exchange下推 / Sort下推



Join重排

- 避免空间膨胀：分组和限制
- 分布式环境特点：稠密树优先

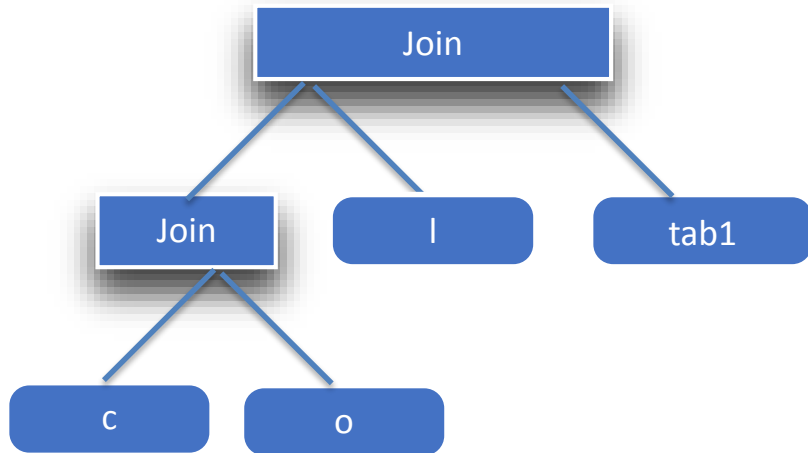


Join合并

- Join重排和合并作为统一的优化过程

(TPC-H Q18)

```
...
from customer c
join orders o on c.c_custkey = o.o_custkey
join lineitem l on o.o_orderkey = l.l_orderkey
join
(
  ...
) tab1
on o.o_orderkey = tab1.l_orderkey
...
```



自动的MapJoin

- MapJoin自动转换
 - 默认打开
 - 保守的策略
- Broadcast on-the-fly



统计数据

- 用途
 - 代价计算 : $\text{Cost} = f_{\text{cost_model}}(\text{Expression}, \text{Statistics})$
 - 应用规则
- 类别
 - 表 : RowCount, FileSize, AvgRowSize, ...
 - 列 : Distinct (NDV), MaxValue / MinValue, AvgColumnSize, ...
 - 复杂 : TopKValues, Histogram, ...
- 收集方式
 - Analyze
 - 自动收集



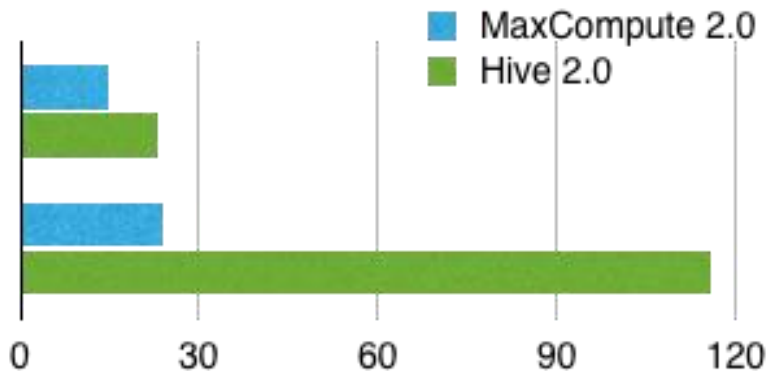
基于LLVM的高效代码生成

- 基于LLVM的高效代码生成
- 向量化执行

part (p_brand <> 'Brand#45'
and p_type not like 'MEDIUM POLISHED%'
and p_size in (49, 14, 23, 45, 19, 3, 36, 9))

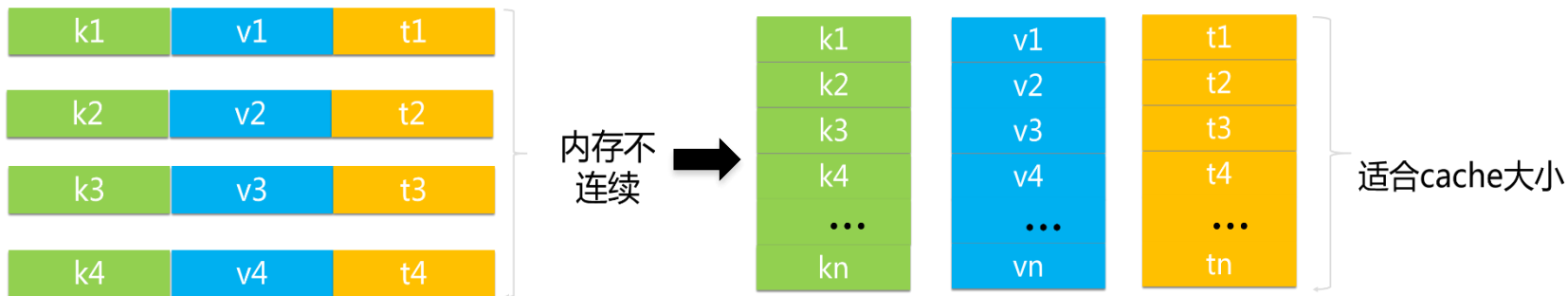
supplier(s_comment like '%Customer%Complaints%')

Table Scan Improvements Based on LLVM
(TPC-H Q16)



向量化执行(SIMD)

- 缓存友好的算法
- 利用现代CPU特性 (SSE/AVX)



Roadmap

- Join reordering增强
- Range partitioning支持
- 自动统计信息收集
- 更多的运行时算法
- Hash Aggregate
- Shuffled Hash Join
- 兼容社区生态系统
- 支持TopKValues统计信息和Skew join
- ...



总结

- MaxCompute 2.0 SQL 是一个重大更新
- 新版本的性能有长足的进步
- 查询优化减少了大部分人工干预自动统计信息收集
- （我们在招人）



2016 The
Computing
Conference
THANKS






2016 杭州·云栖大会
THE COMPUTING CONFERENCE

云栖社区
yq.aliyun.com

Backup Slides

2016
The Computing Conference

主办单位:  杭州

 Alibaba Group
阿里巴巴集团

战略合作伙伴: 



扫码观看大会视频

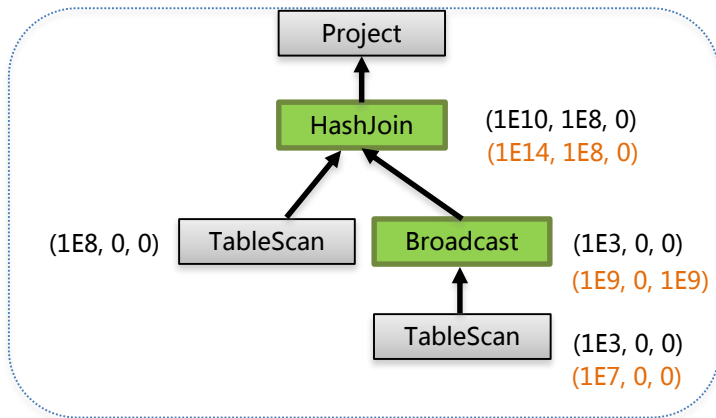
基于代价的优化器

- 代价计算

- 代价：(RowCount, CPU, IO) 三元组
- 代价模型
- 范例 *

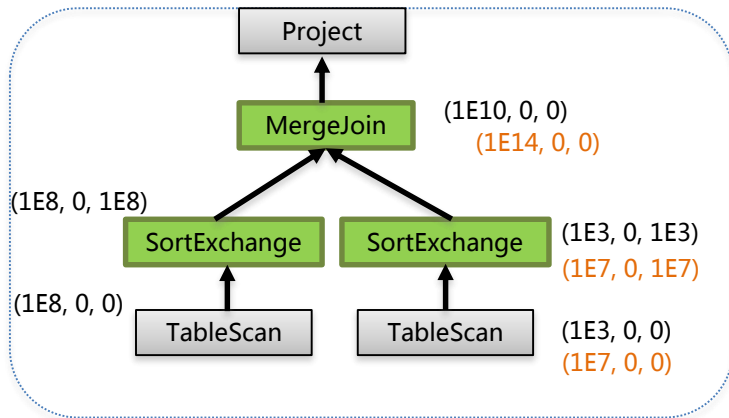
总计：(10100101000, 1E8, 1E5) ✓

(100001110000000, 1E8, 1E5)



总计：(10200002000, 0, 100001000)

(100000220000000, 0, 110000000) ✓



* 专利：分布式数据仓库中一种估计Join计算代价的方法



扫码观看大会视频

Aggregate Rule

- 3种Aggregate
 - 1 pass : SortedAgg
 - 2 pass : HashAgg + SortedAgg
 - 3 pass : Deduplicate + SortedAgg
- + SortedAgg
- Cost based
 - 不再需要skewindata设定

