# 概述

# 优化器

## TDSQL

## PolarDB

## GoldenDB

和单机数据库相比，GoldenDB分布式数据库系统中多了一层全局概念模式到本地概念模式的映射。表数据也被水平切分到多个数据节点。分布式查询优化器的优劣直接影响着业务性能。GoldenDB分布式数据库查询优化器主要朝着如下三个方向努力：

1. 最大程度使数据操作本地化、局部化，减少网络通信的交互次数和交换的数据量，提升数据节点的并行计算是分布式数据库系统优化器致力的重要方向之一；
2. 如何降低数据全局一致性保证的开销，也是分布式数据库系统查询优化努力的方向；
3. 在分布式数据库系统中使用单机数据库的优化手段或者其变形，从而使存储节点承担更多的优化工作，在全局层面仅作少量的启发式优化。

### 影响因素

优化器的优化工作主要体现在计划树的生成上，GoldenDB的查询优化器设计实现主要考虑以下两个方面：

1. 代价模型的选择。

GoldenDB采用分布式系统代价估算模型，考虑节点间传输数据的代价，以减少数据传输的次数和数据量作为查询优化的目标，提高数据节点之间计算的并行度、减少计算节点的计算量。这主要考虑在分布式数据库系统环境中，表结构被水平或垂直拆分到多个数据节点，因此需要考虑语句如何分拆、分片之间数据如何移动、结果如何计算与合并的问题，网络通信开销不可忽视。

1. 考虑数据一致性开销。

在分布式数据库系统中，数据全局一致性机制相较于单机数据库需要更为复杂的控制。因此，如何降低数据全局一致性保证的开销，也是GoldenDB查询优化器的设计要求。

总体来讲，GOldenDB的分布式查询优化器遵循了上述的设计原则，以基于规则的优化为主，基于成本的优化为辅，在提升系统的灵活性的同时控制系统实现的复杂性。优化器内部内置大量的优化规则，通过查询重写的方式进行经验性优化。在优化规则的选择上，重点分析分片剪枝、并行执行、合并下压、条件下推、条件繁殖、排序消除、去重消除、排序下推等。

内置大量的优化规则，对上百个场景进行优化，复杂SQL语句兼容性和处理性能好，同时支持prepare预处理、执行计划缓存、数据集透传等功能，保证数据一致性条件下实现高性能SQL处理。

支持的典型优化包括：

1. 分片剪枝
2. 合并下压
3. 并行执行
4. 条件下推及条件繁殖
5. 排序下推、limit下推等
6. 聚合函数优化

得益于完善的优化器设计，使得GoldenDB对单节点、跨节点的复杂SQL的兼容支持程度很高，包括跨节点SUM、COUNT、AVG等汇聚类操作，跨节点WHERE、FROM等子查询，跨节点JOIN，跨节点GROUP BY、ORDER BY、LIMIT等。这是GoldenDB将Proxy命名为计算节点而非中间件的原因之一，也是其和很多分布式数据库产品中间件的主要区别。

### 合并下压优化

#### 表层次

##### 复制表+复制表

如果dup\_1所在的db分组与dup\_2所在的db分组有公共的交集，则可以合并下发，否则不能合并下发该语句；无需考虑等值链和值链对合并结果的影响。

例如：

dup\_t1分布在g1,g2,g3，dup\_t2分布在g1,g3

结论：合并下发

dup\_t1分布在g1,g3，dup\_t2分布在g2,g4

结论：不可以合并下发

##### 复制表+非复制表

1. 无where条件或者where条件为非分发键

如果复制表所在的db分组要完全覆盖非复制表所在的分组，则可以合并下发，否则不能合并下发该语句。

例如：

dup\_t分布在g1,g2,g3，非复制表分布在g1,g2,g3

结论：可以合并

dup\_t分布在g1,g2，非复制表分布在g1,g3

结论：不可以合并

1. where条件为分发键

如果非复制表经过where值的过滤后能落在单个db上，则判断依据参考复制表+复制表合并依据，否则参考复制表+非复制表且无where条件或where条件为非分发键合并依据。

例如：

dup\_t分布在g1,g3，非复制表分布在g1,g2

where条件：range\_t.a>20and range\_t.a<50

结论：可以合并

dup\_t分布在g1,g3，非复制表分布在g1,g2

where条件：range\_t.a>100 and range\_t.a<250

结论：不可以合并

##### hash表+hash表

1. 无where条件或where条件为非分发键

如果两个hash表所在的db分组一致且关联字段都是两个表的分发键时，则可以合并下发，否则不可合并下发。

1. where条件为分发键

如果通过值链和等值链过滤后有一个表被过滤落在单个db，则判断依据参考复制表+非复制表且where为分发键的合并依据。

如果通过值链和等值链过滤后两个表都落在单个db上，则判断依据参考复制表+复制表合并依据；否则参考hash表+hash表且无where条件或where为非分发键合并依据。

##### range表+range表

1. 无where条件或者where条件非分发键

如果两个range表同时满足如下条件，并且join on关联条件为每个表的分发键时，则可以合并下发，否则不可合并下发：

1. 两表的db分布完全一致
2. 满足条件1后，range1表在每个db上的分布范围也必须和range2在每个db上的分布范围完全一致

例如：

range1分布在g1范围为[-∞,-100)，g2范围为[100,200)

range2分布在g1范围为[-∞,-100)，g2范围为[100,200)

结论：可以合并

range1分布在g1范围为[-∞,-50)，g2范围为[50,100)

range2分布在g1范围为[-∞,-60)，g2范围为[60,100)

结论：不可以合并

1. where条件为分发键

如果两个range表同时满足如下条件，并且join on关联条件为每个表的分发键时，则可以合并下发，否则不能合并下发：

1. 两表的db分布完全一致
2. 满足1条件后，经过where过滤后的range1表的g1和g2范围分布落在range2表原始大范围内，同时经过where过滤后的range2表的g1,g2小范围落在range1的g1,g2原始大范围内

例如：

range1分布在g1范围为[-∞,-50)，g2范围为[50,100)

range2分布在g1范围为[-∞,-60)，g2范围为[60,90)

where条件为：

range1.c>30 and range1.c<40 and range2.b>30 and range2.b<30

结论：可以合并

range1分布在g1范围为[-∞,-50)，g2范围为[50,100)

range2分布在g1范围为[-∞,-60)，g2范围为[60,150)

where条件为：

range1.c>30 and range1.c<40 and range2.b>110 and range2.b<140

结论：不可以合并

##### list表+list表

1. 无where条件或者where条件为非分发键

如果两个list表同时满足如下条件，并且join on关联条件为每个表的分发键时，则可以合并下发，否则不可合并下发：

1. 两表的db分布完全一致
2. 满足条件1后，list1表在某个db上的分布列表值，不能出现在list2表除了g1之外的任何db上

例如：

list1分布在g1范围为(1,6,7)，g2范围为(9,11,12)

list2分布在g1范围为(1,6,8)，g2范围为(9,13)

结论：合并下发

list1分布在g1范围为(1,6,7)，g2范围为(8,11,12)

list2分布在g1范围为(1,6,8)，g2范围为(9,13)

结论：不能合并下发

list1分布在g1范围为(1,6,7)，g2范围为(9,11,12)

list2分布在g1范围为(3,5)，g2范围为(9,13)

结论：合并下发

1. where条件为分发键

如果两个list表同时满足如下条件，并且join on关联条件为每个表的分发键时，则可以合并下发，否则不可合并下发：

1. 两表的db分布完全一致
2. 满足条件1后，经过where过滤后的list1表的g1和g2小范围分别落在list2表的g1,g2原始大范围内，同时经过where过滤后的list2表的g1,g2小范围分别落在list1表的g1,g2原始大范围内

例如：

list1分布在g1范围为(1,6,7)，g2范围为(9,11,12)

list2分布在g1范围为(1,6,8)，g2范围为(9,13)

where条件为：(list1.c=1 and list2.c=6) or (list1.c=11 or list2.b=9)

list1过滤后范围：g1范围(1)，g2范围(11)

list2过滤后范围：g1范围(6)，g2范围(9)

结论：合并下发

list1分布在g1范围为(1,6,7)，g2范围为(8,11,12)

list2分布在g1范围为(1,6,8)，g2范围为(9,11)

where条件为：(list1.c=1 and list2.b=6)

list1过滤后范围：g1范围(1)

list2过滤后范围：g1范围(6)

结论：合并下发

list1分布在g1范围为(1,6,7)，g2范围为(8,11,12)

list2分布在g1范围为(1,6,8)，g2范围为(9,11)

where条件为：(list1.c=1 and list2.b=8)

list1过滤后范围：g1范围(1)

list2过滤后范围：g1范围(8)

结论：不可以合并下发

#### 主子查询

##### FROM子查询

不能合并的场景：

1. 如果FROM子查询本身不是SQLNode或MSQLNode

代表子查询的JoinNode或UnionNode，不能合并入主查询，因为本身就不可合并

1. 同时满足：多db分布（handle\_on=3）&（有汇聚函数/distinct）&（非复制表）&（分发键不在group中）

如果非复制表多db分布且分发键不在group by中，则汇聚函数必须在proxy层做聚合和distinct计算

1. 如果FROM子查询中有不能合并的子查询时
2. 同时满足：FROM子查询多db分布且有limit

limit需要proxy计算

##### WHERE子查询

不能合并的场景：

1. where子查询为JoinNode、UnionNode
2. CR场景，where子查询不能与主查询合并

CR场景涉及gtid列，则where子查询中的gtid无法处理，比如：where t1.a in (select t2.a,t2.gtid ...)

1. Where子查询为非复制表多db分布，且有聚合函数
2. 主查询中标为复制属性，子查询表为非复制属性，分布于多个db且查询字段为非分发键时，主子查询不能合并下发（不管查询字段是否为分发键，都不能合并）

###### IN/NOT IN子查询

1. SPJ查询语句

合并下发前提：

1. 涉及的所有表都是hash分布
2. 主子查询能直接使用分发键关联
3. 多表的分发类型需要一致
4. 分发键需要NOT NULL限制

合并下发原理：

在分发键类型一致的情况下，通过唯一的hash算法可以将相同值统一的发布到相同的数据节点上。这样可以保证not in子查询数据分布的一致性，因此可以考虑合并下发。

1. 非SPJ查询语句
2. 子查询有order by，有group by+分发键，子查询投影中为分发键且有distinct，having子句

上述对合并下发没有影响。

1. 子查询group by中没有分发键：不能合并

汇聚函数计算需要在所有db数据上进行，如果没有汇聚函数情况下，group by也会隐藏部分分发键值。

1. 子查询有limit：不能合并

limit需要在所有db数据上进行操作。

1. 子查询投影中有单纯的汇聚函数且group by中包含子查询表的分发键：不能合并

在子查询投影中有汇聚函数，即使有group by+分发键也不能合并。

###### EXISTS/NOT EXISTS子查询

1. 非关联子查询：不能合并

exist/not exists存在和不存在判断是对于子查询表的所有数据生效。

1. SPJ关联子查询

合并下发前提：

1. 涉及的所有表都是hash分布
2. 主子查询能够直接使用分发键等值关联
3. 多表的分发键类型需要一致
4. 分发键需要NOT NULL限制

合并下发原理：在分发键类型一致的情况下，通过唯一的hash算法可以将等值统一的发布到相同的数据节点上。这样可以保证关联的exists子查询数据分布的一致性。因此可以考虑合并下发。

1. 非SPJ关联子查询

1）子查询有order by，有group by+分发键，having子句

上述对合并下发没有影响。

2）子查询group by中没有分发键：不能合并

汇聚函数计算需要在所有db数据上进行，如果没有汇聚函数情况下，group by也会隐藏部分分发键值。

3）子查询有limit：不能合并

limit需要在所有db数据上进行操作。

###### IN/NOT IN item\_row子查询

1、SPJ查询语句

合并下发前提：

1）涉及的所有表都是hash分布

2）主子查询能直接使用分发键关联

3）多表的分发类型需要一致

4）分发键需要NOT NULL限制

合并下发原理：

在分发键类型一致的情况下，通过唯一的hash算法可以将相同值统一的发布到相同的数据节点上。这样可以保证not in子查询数据分布的一致性，因此可以考虑合并下发。

2、非SPJ查询语句

1）子查询有order by，有group by+分发键，子查询投影中为分发键且有distinct，having子句

上述对合并下发没有影响。

2）子查询group by中没有分发键：不能合并

汇聚函数计算需要在所有db数据上进行，如果没有汇聚函数情况下，group by也会隐藏部分分发键值。

3）子查询有limit：不能合并

limit需要在所有db数据上进行操作。

4）子查询的group by子句中有分发键：与SPJ查询合并前提一致，同时需要子查询的投影与分发键相关联，只有这样才能保证最终结果没有重复数据

在子查询投影中有汇聚函数，即使有group by+分发键也不能合并。

#### UNION

1. 单DB场景

CR：

select1 UNION select2时，如果左右表的个数不一致时不能合并下发，原因是每张表都要添加gtid。

select1 UNION select2时，如果左右表的个数一致时可以进一步判断是否合并下发。

UR：

不论左右select中标的个数是否一致，都需要进一步判断是否可以合并下发。

1. 多DB场景

CR：

如果左右select中表的个数不一致，则不能合并下发。

如果左右select中表的个数一致，需要进一步判断是否可以合并下发。

UR：

select1 UNION select2同时满足如下条件才可以考虑合并下发：

1. 左右select中没有group by、没有聚合函数同时没有distinct
2. 没有使用union distinct

### SQL引擎优化

#### 条件繁殖

条件繁殖是指优化器对已知条件进行推断，从而衍生出其他条件进行改造，缩小数据检索的范围。繁殖后的条件，或推入基表、或下压到数据节点执行。

SQL示例1：

select T1.col,T1.col1 from T1 join T2 on T1.col=T2.col where T1.col>100 UR;

优化后执行计划示例：

优化器会识别出本例中的条件传递从而推断出T2.col>100，并将该条件推入基表。语句被重写：

select T1.col,T1.col1 from {T1 where col>100} join {T2 where col>100} on T1.col=T2.col UR;

注：在等价语句改造的时候，也可以利用ON条件等做等值的传递。

SQL示例2：

select T1.col, T1.col1 from T1 join T2 on T1.col=T2.col where T1.col1 in (100,200) and T2.col1=T1.col1 UR;

优化器执行计划示例：

优化器会识别出本例中的条件传递从而推出T2.col1 in (100,200)，并将该条件推入基表。语句被重写为：

select T1.col, T1.col1 from {T1 where col1 in (100,200)} join {T2 where col1 in (100,200)} von T1.col=T2.col UR;

#### 并行执行

并行执行是指执行计划在各个分区间进行并行执行，从而提升执行效率。

当SQL查询在分区剪裁后，仍然涉及多个分区时，会生成一个分布式执行计划，该分布式计划会被调度到分区所在不同机器上进行执行。GoldenDB在判断语句需要下发到多个分区时，会将语句拆分成多个同时下发到对应的节点并行执行。

SQL示例：

假定T1、T2在g1,g2,g3三个分区上：

select T1.col,T1.col1 from T1 join T2 on T1.col=T2.col where T1.col1=200 UR;

优化后执行计划，语句被拆分如下：

select T2.col from T2 order by T2.col ASC;

select T1.col,T1.col1 from T1 where T1.col1=200 order by T1.col ASC;

同时下发到对应的三个节点上并行执行后，将结果汇总到proxy层做join。

#### AVG优化

在GoldenDB中，AVG被自动重写成SUM和COUNT两个计算，在每个数据节点上，只返回本数据节点的SUM、COUNT；在计算节点层面，再对各数据节点返回的SUM、COUNT进行累计，然后再用SUM/COUNT得到AVG的最终结果。

SQL示例：

select avg(col) from T where col>100 UR;

优化后执行计划示例：

select sum(col),count(col) from T where col>100;

#### WHERE条件下推

在分布式数据库系统实现中，为了尽量减少数据节点向计算节点移动的数据量，系统被设计为尽可能将where条件下推到数据节点。

SQL示例：

假定T为单分发键表，且分发键为col：

select col,col1 from T where col=100 and col1>10 UR;

优化后执行计划示例：

根据where子句可以将查询数据定位到某（几）个GROUP，执行时将语句直接下发到对应的group节点上执行：

select col,col1 from T where col=100 and col1>10;

#### order by下推

GoldenDB对于排序处理通常会优先考虑推入数据节点完成。利用数据节点的计算能力并行完成排序操作；涉及结果合并的，计算节点再对有序数据集进行合并排序。

SQL示例：

select col,col1 from T where col1>100 order by col UR;

优化后执行计划示例：

将语句下发到各个节点上并行执行，并在proxy层汇总结果，如果下发为多节点，则需要执行sort merge排序操作。

select col,col1 from T where col1>100 order by col;

#### distinct下推

GoldenDB中，遇到不能合并下发的SQL语句，如果其中含有distinct，则计算节点在拆分语句时，会考虑将distinct下推入数据节点执行。以减少从数据节点提取到计算节点的数据量。

UR场景下，如果查询数据分布在同一个节点上或者select list为分发键的情况下，将distinct下推入节点执行，汇总结果不需要在proxy层再做distinct。

SQL示例：

假定T为多分发键表，且分发键为col1,col1：

select distinct col,col1 from T where col>100 and col1=20 UR;

优化后执行计划示例：

select distinct col,col1 from T where col>100 and col1=20;

#### limit下推

在GoldenDB中，limit下推的主要目的是在需要计算节点进一步计算的场景下，尽量减少从数据节点提取到计算节点的数据量。

优化原则：

1. SQL语句能下发到一个db group执行的，limit子句不用调整
2. SQL语句下发到多个db group执行的，需要在proxy层汇总数据，做limit操作：
3. SQL语句能下发，但是需要下发到多个db group执行的，limit子句需要调整，调整格式如下：limit x,y --> limit 0,x+y
4. SQL语句不能下发，需要把数据拉到db proxy层计算的，limit子句不变

SQL示例：

查询数据分布在多个节点上，且查询语句能下发：

select col,col1 from T where col1>100 limit 2,2 UR;

优化后执行计划：

select col,col1 from T where col1>100 limit 0,4;

#### 常数折叠

在GoldenDB中，为了减少对确定值的反复计算而先进行计算的优化方法。此过程一般发生在S（Select）F（From）W（Where）中SW阶段。

SQL示例：

假定T为单分发键表，且分发键为col：

select col,col1 from T where col=50+50 and col1>10 UR;

优化后执行假话示例：

在此查询中，50+50会被先计算橙100。查询重写后，条件变为where col=100 and col1>10。根据重写后的where条件可以将查询语句定位到某（几）个group上执行，执行时下发原始条件。

select col,col1 from T where col=50+50 and col>10;

#### 非逻辑优化

在GoldenDB中，会针对NOT运算进行处理，通常是将其下推，将表达式整体取反变为表达式分量补集的运算。

|  |  |
| --- | --- |
| 处理前 | 处理后 |
| NOT (col!=5) | col=5 |
| NOT(col1<=4 OR col2>0) | col1>4 AND col2<=0 |
| NOT(col1<=4 AND col2>0) | col1>4 OR col2<=0 |

经过变换，可以减少一次逻辑运算并在一定条件下使范围扫描可用。

#### 死代码消除

GoldenDB分布式优化器中的处理逻辑通过判断出为恒指或者逻辑冗余的条件，然后在运行时减少不必要的逻辑判断，从而提升执行效率。

SQL示例：

假定T为range表，col为单分发键：

select col,col1 from T where col>0 and col > 200 UR;

优化后执行计划：

此例子中，根据where条件确定下发group（col>200所在group）：

select col,col1 from T where col>0 and col>200;

SQL示例：

select col,sol1 from T where col<0 and col>200 UR;

优化后执行计划：

此例中，根据where条件获取下发group为0，选择一个group下发执行：

select col,col1 from T where col<0 and col>200;

#### 合并下发

GoldenDB的分布式优化，很重要的一个努力方向就是尽量利用数据节点的计算能力进行计算，避免不必要的从数据节点向计算节点的数据移动，并减少和数据节点交互的次数。计算节点分析语句后，尽量把能够一起执行的语句下发到数据节点。

注：这里涉及到表层次合并、主子查询合并以及JOIN等合并下发规则的判断。

SQL示例：

假定T1和T2分发属性相同：range表、分发键为col、分布在g1,g2,g3节点上，将“条件繁殖”部分的示例语句改写如下：

select T1.col,T1.col1 from T1 join T2 on T1.col=T2.col where T1.col=200 UR;

见“where条件下推”部分，根据where子句可以将查询数据定位到某个group上，则可以直接将语句下发到group节点上执行。

优化后执行计划：

select T1.col,T1.col1 from T1 join T2 on T1.col=T2.col where T1.col=200;

### 执行器

## OceanBase

## TiDB