链滴

记一次数据类型不同导致的 sql join 异常

作者：bivana

原文链接：https://ld246.com/article/1564719517719

来源网站：链滴

#奇葩现象

我们知道，常用的语句

```
select count(distinct a.id) as id1
,count(distinct b.id) as id2
from table_a a
left  join table_b b
on a.id=b.id
```

结果逻辑上id1必定大于等于id2，因为a表为主表，且b表为被关联表，包含条件a.id=b.id，也就是说表有的id a表也都存在。

但是今天在spark-sql 场景下，本人碰到了id2的值大于id1的值的奇葩现象，一开始，百思不得其解，后来经过数据抽样排查，发现是数据类型不一致导致的锅，具体来说，就是int 的 1和 001的字符类型关联上了

现象复现：

```
--创建int主键类型的测试表
drop  table  if  exists temp.int_primary;
create  table  if  not  exists temp.int_primary(
id int comment 'int 型id'
);

--写入int类型的主键值
insert  into  table temp.int_primary
select id
from
  ( select 1  as id ) a ;

--数据查看
select * from temp.int_primary;
--结果
id
---
1


--创建字符串类型的结果表
drop  table  if  exists temp.string_primary;
create  table  if  not  exists temp.string_primary(
id string comment 'string 型id'
)
;

--写入字符串类型表数据
insert  into  table temp.string_primary
select id
from
  (select '1'  as id
   union
   select '01'  as id
   union
```

```
   select '001'  as id
) a ;

--查看数据结果
select * from temp.string_primary;
--结果
id
---
1
01
001

--见证奇迹的时刻，使得被关联表id更多
select count(distinct a.id) as id1
,count(distinct b.id) as id2
from temp.int_primary a
left join temp.string_primary b
on a.id=b.id

--结果
id1 id2
----
1    3
```

如果按以上操作，必定能复现次奇怪现象。

原因分析

其实，看了现象复现，对于产生原因，笔者相信大部分人都会有响应的猜测了。

经过本人测试，发现spark-sql进行sql解析时，对于 a left join b或a right join b这种操作时，如果联的字段类型不一致，会以主表的字段类型为基础，将被关联表的字段类型转为主表的字段类型进行联，如果被关联表的字段类型无法转换（如字符串‘sdfsd’无法转换为数字），那么会被当成null处。

在以上的例子中,b表的 01和 001字符串，被转成int型后为1，和a表关联上了，所以关联后的结果有条。

但是在count(distinct b.id)这个语句时，获取的又是未经处理的字符串类型，所以统计出来的3，而非.

至此，奇怪的现象已经解释完毕，spark-sql在将sql转换为spark程序时，对部分字段类型进行转换，部分未转换，造成了结果的不一致。