Hive-数组类型字段取值为[]或NULL使用注意事项

描述

数组类型下可能出现NULLl和[]的情况

- NULL的情况一般为outer join造成的
- []的情况一般为人工处理产生的

思考

既然存在[]或NULL的情况，在使用上有什么坑吗？主要担心点为explode拆数组的时候会出现什么问题吗？

测试结论

（1）explode

- explode在拆数组类型无论是[]还是null都不会产生问题，拆解出来都是为NULL。
- 但是在配合使用lateral view的时候可能会出现问题，lateral view在连接数据的时候采用的是笛卡尔积的形式，不能连接上为NULL的数据，导致结果会缺少数组为[]或NULL的数据。
- lateral view outer可以将所有数据进行连接，包括数组为[]或NULL的情况，连接后的数据为NULL。
- 在使用过程中，如果不能保证该数组字段中没有[]或NULL出现，推荐使用lateral view outer，避免丢失数据。

（2）size

- 当该数组字段取值为[]时，size=0
- 当该数组字段取值为NULL时，size=-1

（3）arr_contains

- 当该数据字段取值为[]时，返回false
- 当改数组字段取值为NULL时，返回NULL

（3）取值

- 当该数据字段取值为[]或NULL时，取到任意位置值都为NULL

测试过程

测试代码：

```
select user_id
      ,ks_order_list --
      ,size(ks_order_list) as list_size --
      ,ks_order_list[0] as index_zero --
      ,array_contains(ks_order_list,"123") as contains
      ,lvtb.ks_order as ks_order --
from
(
        --union alllimitlimit
        --ks_order_listnull
    select *
    from
    (
        select user_id
              ,ks_order_list
        from kscdm.dwd_ks_csm_play_live_hi
        where p_date = '20200506'
        and ks_order_list is null
        limit 1
    ) a

    union all

        --ks_order_list[]
    select *
    from(
        select user_id
              ,ks_order_list
        from kscdm.dwd_ks_csm_play_live_hi
        where p_date = '20200506'
        and p_hour = '18'
        and size(ks_order_list) = 0
        limit 2
    ) b

    union all

        --ks_order_list
    select *
    from(
        select user_id
              ,ks_order_list
        from kscdm.dwd_ks_csm_play_live_hi
        where p_date = '20200506'
        and p_hour = '18'
        and size(ks_order_list) between 2 and 3
        limit 2
    ) c
) t
--lateral view outer
lateral view outer explode(ks_order_list) lvtb as ks_order
--lateral view
lateral view  explode(ks_order_list) lvtb as ks_order
```

使用lateral view outer

user_id	ks_order_list	list_size	index_zero	contains	ks_order
362913750	NULL	-1	NULL	NULL	NULL
219991974	[]	0	NULL	false	NULL
889814696	[]	0	NULL	false	NULL
1856643977	["pkoi:51","pkoi:1025"]	2	pkoi:51	false	pkoi:51
1856643977	["pkoi:51","pkoi:1025"]	2	pkoi:51	false	pkoi:1025

736494488	["S-161BE8F694E8","S-161BE8F694E8"]	2	S-161BE8F694E8	false	S-161BE8F694E8
736494488	["S-161BE8F694E8","S-161BE8F694E8"]	2	S-161BE8F694E8	false	S-161BE8F694E8

使用lateral view

user_id	ks_order_list	list_size	index_zero	contains	ks_order
1856643977	["pkoi:51","pkoi:1025"]	2	pkoi:51	false	pkoi:51
1856643977	["pkoi:51","pkoi:1025"]	2	pkoi:51	false	pkoi:1025
736494488	["S-161BE8F694E8","S-161BE8F694E8"]	2	S-161BE8F694E8	false	S-161BE8F694E8
736494488	["S-161BE8F694E8","S-161BE8F694E8"]	2	S-161BE8F694E8	false	S-161BE8F694E8

统一格式

是否有必要将NULL和[]统一？

Hive中insert语句必须列数匹配，不支持不写入，没有值的列必须使用null占位，无数据时自动补。

Hive中NULL值底层默认存储为 ‘\N’ ，可以通过alter table name SET SERDEPROPERTIES(‘serialization.null.format’ = ‘\N’)来制定。

如果数据内容为 ‘\N’ ，显示出来也为NULL；通过也可以通过 = ‘\N’ 来代替is null。

```
select user_id
      ,ks_order_list
      ,case
          when ks_order_list is null then array()
          else ks_order_list
       end as format1
      ,case
          when size(ks_order_list) = 0 then null
          else ks_order_list
       end as format2
from
(
    select *
    from
    (
        select user_id
              ,ks_order_list
        from kscdm.dwd_ks_csm_play_live_hi
        where p_date = '20200506'
        and ks_order_list is null
        limit 1
    ) a

    union all

    select *
    from(
        select user_id
              ,ks_order_list
        from kscdm.dwd_ks_csm_play_live_hi
        where p_date = '20200506'
        and p_hour = '18'
        and size(ks_order_list) = 0
        limit 2
    ) b

    union all

    select *
    from(
        select user_id
              ,ks_order_list
        from kscdm.dwd_ks_csm_play_live_hi
        where p_date = '20200506'
        and p_hour = '18'
        and size(ks_order_list) between 2 and 3
        limit 2
    ) c
) t
```

统一结果：

user_id	ks_order_list	format1	format2
745352495	NULL	[]	NULL
219991974	[]	[]	NULL
1207219075	[]	[]	NULL
1491110766	["S-B23F70962E4F","1588739304905000_1132067172_1491110766_"]	["S-B23F70962E4F","1588739304905000_1132067172_1491110766_"]	["S-B23F70962E4F","1588739304905000_1132067172_1491110766_"]
550860007	["HSF-VPP4VHMJJ8CB","S-997067D89C39"]	["HSF-VPP4VHMJJ8CB","S-997067D89C39"]	["HSF-VPP4VHMJJ8CB","S-997067D89C39"]