Hive-记：热点数据导致数据倾斜问题

背景：

任务：kscdm.dws_ks_csm_prod_user_photo_funnel_1d

该任务在20200701日执行时间为4h,怀疑出现数据倾斜。初步怀疑是Join导致的数据倾斜。

```
416             where  p_date= '{{ds_nodash}}'
417             and    photo_id > 0
418             and    negative_type in ('report_photo', 'reduce_photo')
419             group by nvl(user_id,0)
420                     ,coalesce(product,'UNKNOWN')
421                     ,negative_type
422             )t4
423       group by user_id
424               ,product
425     )a
426 left join (
427       select  nvl(user_id,0) as user_id
428               ,product
429               ,browse_type
430               ,feed_model
431               ,country_region
432               ,user_active_degree
433               ,device_brand
434               ,device_model
435               ,platform
436               ,app_version
437               ,app_minor_version
438               ,app_major_version
439               ,country_name
440               ,province_name
441               ,city_name
442       from    kscdm.dws_ks_usr_prod_user_active_1d
443       where   p_date = '{{ds_nodash}}'
444       )b
445 on          a.user_id=b.user_id
446 and         a.product=b.product
447 group by nvl(a.user_id,0)
448          ,a.product
449 ;
450 alter table ${db['kscdm']}.dws_ks_csm_prod_user_photo_funnel_1d add if not exists p
```

该任务是核心报表的上游，因为其数据倾斜导致后续报表进程delay，急需优化。

排查：

（1）首先，查看该天的任务运行日志。可以看到，明显stage-1在执行过程中reduce端数据倾斜。

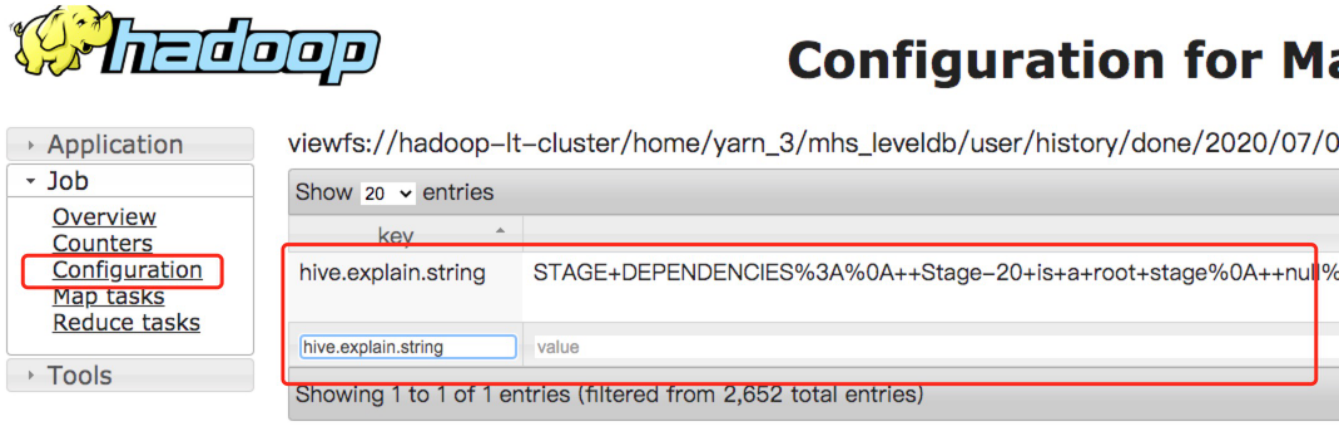
```
[2020-07-01T05:19:04.124]INFO  : 2020-07-01 05:19:04,043 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2176856.31 sec
[2020-07-01T05:20:09.149]INFO  : 2020-07-01 05:20:04,840 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2176901.8 sec
[2020-07-01T05:21:09.173]INFO  : 2020-07-01 05:21:05,443 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2176949.09 sec
[2020-07-01T05:22:09.197]INFO  : 2020-07-01 05:22:06,430 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2177002.79 sec
[2020-07-01T05:23:09.220]INFO  : 2020-07-01 05:23:07,118 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2177052.24 sec
[2020-07-01T05:24:09.244]INFO  : 2020-07-01 05:24:07,842 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2177104.22 sec
[2020-07-01T05:25:09.263]INFO  : 2020-07-01 05:25:08,158 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2177145.93 sec
[2020-07-01T05:26:09.285]INFO  : 2020-07-01 05:26:08,900 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2177200.27 sec
[2020-07-01T05:27:14.310]INFO  : 2020-07-01 05:27:09,486 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2177249.29 sec
[2020-07-01T05:28:14.436]INFO  : 2020-07-01 05:28:10,548 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2177303.19 sec
[2020-07-01T05:29:14.461]INFO  : 2020-07-01 05:29:11,313 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2177347.39 sec
[2020-07-01T05:30:14.569]INFO  : 2020-07-01 05:30:12,146 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2177397.07 sec
[2020-07-01T05:31:14.653]INFO  : 2020-07-01 05:31:12,854 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2177442.84 sec
[2020-07-01T05:32:14.679]INFO  : 2020-07-01 05:32:13,781 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2177485.92 sec
[2020-07-01T05:33:14.708]INFO  : 2020-07-01 05:33:14,324 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2177536.75 sec
[2020-07-01T05:34:19.736]INFO  : 2020-07-01 05:34:15,170 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2177599.4 sec
[2020-07-01T05:35:19.760]INFO  : 2020-07-01 05:35:15,875 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2177654.96 sec
[2020-07-01T05:36:19.868]INFO  : 2020-07-01 05:36:16,900 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2177719.96 sec
[2020-07-01T05:37:19.915]INFO  : 2020-07-01 05:37:17,410 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2177773.77 sec
[2020-07-01T05:38:19.938]INFO  : 2020-07-01 05:38:18,324 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2177836.32 sec
[2020-07-01T05:39:19.963]INFO  : 2020-07-01 05:39:18,935 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2177891.21 sec
[2020-07-01T05:40:20.174]INFO  : 2020-07-01 05:40:19,591 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2177949.88 sec
[2020-07-01T05:41:20.672]INFO  : 2020-07-01 05:41:20,306 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2178004.95 sec
[2020-07-01T05:42:25.381]INFO  : 2020-07-01 05:42:21,138 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2178063.17 sec
[2020-07-01T05:43:25.509]INFO  : 2020-07-01 05:43:21,541 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2178119.62 sec
[2020-07-01T05:44:25.563]INFO  : 2020-07-01 05:44:22,529 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2178181.08 sec
[2020-07-01T05:45:25.588]INFO  : 2020-07-01 05:45:22,953 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2178231.92 sec
[2020-07-01T05:46:25.700]INFO  : 2020-07-01 05:46:23,941 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2178296.97 sec
[2020-07-01T05:47:25.728]INFO  : 2020-07-01 05:47:24,670 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2178345.56 sec
[2020-07-01T05:48:25.755]INFO  : 2020-07-01 05:48:25,666 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2178408.88 sec
[2020-07-01T05:49:30.782]INFO  : 2020-07-01 05:49:26,426 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2178466.62 sec
[2020-07-01T05:50:30.806]INFO  : 2020-07-01 05:50:27,348 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2178518.66 sec
[2020-07-01T05:51:30.829]INFO  : 2020-07-01 05:51:30,663 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2178585.93 sec
[2020-07-01T05:52:35.990]INFO  : 2020-07-01 05:52:31,408 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2178637.76 sec
[2020-07-01T05:53:36.014]INFO  : 2020-07-01 05:53:32,227 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2178700.23 sec
[2020-07-01T05:54:36.036]INFO  : 2020-07-01 05:54:32,651 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2178753.96 sec
[2020-07-01T05:55:36.205]INFO  : 2020-07-01 05:55:33,405 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2178815.55 sec
[2020-07-01T05:56:36.239]INFO  : 2020-07-01 05:56:33,824 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2178869.64 sec
[2020-07-01T05:57:36.263]INFO  : 2020-07-01 05:57:34,451 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2178932.79 sec
[2020-07-01T05:58:36.289]INFO  : 2020-07-01 05:58:34,956 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2178985.65 sec
[2020-07-01T05:59:36.312]INFO  : 2020-07-01 05:59:35,560 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2179044.21 sec
[2020-07-01T06:00:36.335]INFO  : 2020-07-01 06:00:36,149 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2179093.1 sec
[2020-07-01T06:01:41.369]INFO  : 2020-07-01 06:01:36,834 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2179157.81 sec
[2020-07-01T06:02:41.392]INFO  : 2020-07-01 06:02:37,341 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2179212.1 sec
[2020-07-01T06:03:41.496]INFO  : 2020-07-01 06:03:38,188 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2179273.42 sec
[2020-07-01T06:04:41.511]INFO  : 2020-07-01 06:04:38,628 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2179327.09 sec
[2020-07-01T06:05:41.545]INFO  : 2020-07-01 06:05:39,833 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2179391.85 sec
[2020-07-01T06:06:41.561]INFO  : 2020-07-01 06:06:40,290 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2179445.85 sec
[2020-07-01T06:07:41.579]INFO  : 2020-07-01 06:07:41,200 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2179510.32 sec
[2020-07-01T06:08:46.603]INFO  : 2020-07-01 06:08:41,842 Stage-1(job_1592534694575_6378674) map = 100%,  reduce = 99%; (pending maps/total maps : 0/24775, pending reduces/total reduces : 0/5120), Cumulative CPU 2179566.29 sec
```

（2）查看执行计划，定位stage-1的SQL语句。

1）explain SQL。执行计划是预估的，实际执行可能不走这个执行计划，不太准确。

2）找到该任务ID，点到该任务的YARN日志界面，在Job的Configuration中查找hive.explain.string获取执行计划。该执行计划是编码后的，需要url解码后才能正常查看。

```
Starting Job = job_1592534694575_6378674(Stage-1), Tracking URL = http://10.48.56.8:8088/proxy/application_1592534694575_6378674/
Kill Command = /home/hadoop/software/hadoop/bin/hadoop job  -kill job_1592534694575_6378674
```



找到发生数据倾斜问题的SQL。可以看到SQL中并没有Join操作，所以不是Join导致的数据倾斜。自然想到是不是user_id=0或user_id is null热点数据导致的数据倾斜。

```sql
-- 曝光指标
select  user_id
        ,product
        ,sum(show_cnt)                          as show_cnt
        ,count(if(show_cnt>0,1,null))           as show_photo_num
        ,0                                      as play_cnt
        ,0                                      as play_photo_num
        ,0                                      as play_duration
        ,0                                      as complete_play_cnt
        ,0                                      as complete_play_photo_num
        ,0                                      as valid_play_cnt
        ,0                                      as play_progress
        ,0                                      as comment_stay_duration
        ,0                                      as like_cnt
        ,0                                      as like_photo_num
        ,0                                      as click_like_cnt
        ,0                                      as double_click_cnt
        ,0                                      as cancel_like_cnt
        ,0                                      as cancel_like_photo_num
        ,0                                      as comment_cnt
        ,0                                      as comment_photo_num
        ,0                                      as direct_comment_cnt
        ,0                                      as reply_comment_cnt
        ,0                                      as delete_comment_cnt
        ,0                                      as delete_comment_photo_num
        ,0                                      as comment_like_cnt
        ,0                                      as comment_like_photo_num
        ,0                                      as follow_cnt
        ,0                                      as follow_photo_num
        ,0                                      as cancel_follow_cnt
        ,0                                      as cancel_follow_photo_num
        ,0                                      as share_cnt
        ,0                                      as share_photo_num
        ,0                                      as download_cnt
        ,0                                      as download_photo_num
        ,0                                      as report_cnt
        ,0                                      as report_photo_num
        ,0                                      as reduce_similar_cnt
        ,0                                      as reduce_similar_photo_num
from(
    select  nvl(user_id,0) as user_id
            ,product
            ,photo_id
            ,count(1) show_cnt
    from    ks_dw_fact.dw_fact_client_photo_show_di
    where   p_date= '{{ds_nodash}}'
    and     default.is_prod_user(user_id) = true
    and     default.is_prod_device(device_id) = true
    and     photo_id>0
    and     nvl(user_id,0) > 0
    group by nvl(user_id,0)
             ,product
             ,photo_id
    )t1
group by user_id
         ,product
```

（3）筛选热点数据。从问题表中作简单count，查看一下user_id=0 or user_id is null的数据量。可以看到user_id=0 or user_id is null的数据量很大，导致倾斜。

```
select count(1)
      ,count(if(nvl(user_id,0) = 0, 1 ,null))
      ,count(if(nvl(user_id,0) > 0, 1 ,null))
from    ks_dw_fact.dw_fact_client_photo_show_di
where  p_date= '20200701'
and    default.is_prod_user(user_id) = true
and    default.is_prod_device(device_id) = true
and    photo_id>0
```

	count(1)
total	139673227549
nvl(user_id,0) = 0	3140858221
nvl(user_id,0)! = 0	136532369328

可以到虽然nvl(user_id,0) = 0的条数占比不高，但是依旧有31亿条，试想一下一个reduce处理的话有多艰难。

处理：

（1）热点数据分开处理 union all连接。虽然将热点数据分到一个新的mr中处理，但是因为数据量很大效果不是很明显。

（2）设置hive.groupby.skewindata=true 也可以解决，但是效果同样不是很明显。

如果发生了数据倾斜就可以通过它来进行负载均衡。

该参数当选项设定为 true，生成的查询计划会有两个 MR Job。

第一个 MR Job 中，Map 的输出结果集合会随机分布到 Reduce 中，每个 Reduce 做部分聚合操作，并输出结果，这样处理的结果是相同的Key 有可能被分发到不同的 Reduce 中，从而达到负载均衡的目的；

第二个 MR Job 再根据预处理的数据结果按照Key 分布到 Reduce 中（这个过程是按照key的hash值进行分区的，不同于mr job1的随机分配，这次可以保证相同的Key 被分布到同一个 Reduce 中）

最后完成最终的聚合操作。

注意：（1）和（2）效果不是很明显的关键在于

该操作主要是对photo的distinct的计数，虽然写成了group by + count的形式，因为数据量比较大，性能依旧不是很好。

（3）业务切割。原本user_id = 0 or use_id is null的曝光数据是未登录用户产生的，花大量时间去处理这种使用价值不大的数据不划算，在其他业务中甚至都不统计未登录用户产生的指标。

仅统计nvl(user_id,0)! = 0的曝光数据，nvl(user_id,0)= 0的数据可以直接归0.

结果：

原本任务运行4h，经过优化后仅需要运行1.5h。

后续：

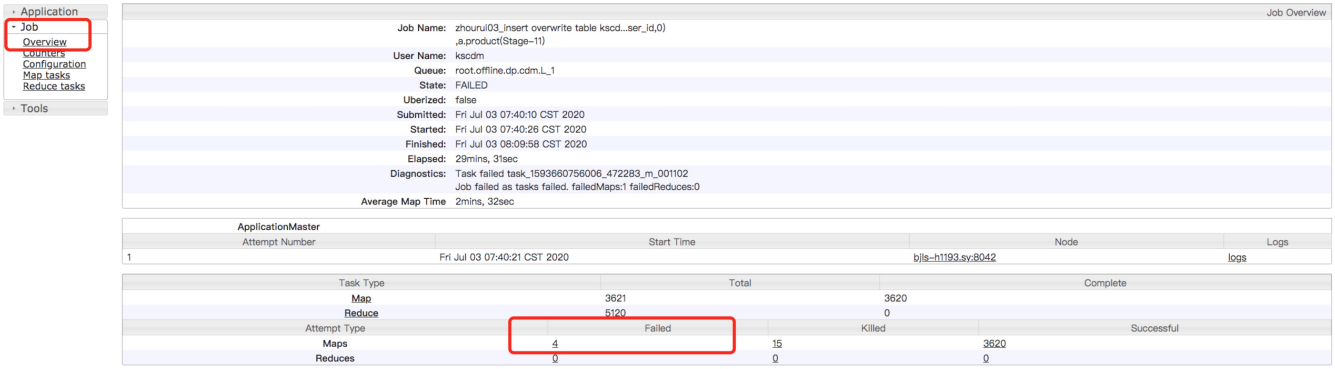
在20200702线上运行该任务的时候，任务报错。

经过排查不存在线上和线下环境不一致情况，是数据本身造成的任务报错。

任务报错为：Hive Task执行失败。

```
|]================================================================================
|]*       【错误解决与优化建议】
|]错误类型:       Hive Task执行失败
|]错误关键字:     Timed out after
|]可能原因:       task超时
|]解决方案:
|]1. 首先请确认是否GC导致超时—登入"FAILED TASK URLs"下的连接，查看LOGS里的jstat.log，如果FGCT超过5min，说明GC严重，请设大内存重试：set mapreduce.map.memory.mb=6144 或者 set mapreduce.reduce.memory.mb=6144;
|]2. 如有join，请参考WIKI调整参数：https://wiki.corp.kuaishou.com/pages/viewpage.action?pageId=80484361#Hue用户使用文档-REDUCETASKTIMEOUT导致作业失败
|]================================================================================
|]*       【详细错误信息】
|]FAILED: Execution Error, return code 2 from org.apache.hadoop.hive.ql.exec.mr.MapRedTask
|]-----
|]Task with the most failures(4):
|]-----
|]Task ID:
|]  task_1593660756006_472283_m_001102
|]FAILED TASK URLs:
|]  http://10.36.255.86:19899/jobhistory/attempts/job_1593660756006_472283/m/FAILED
|]KILLED TASK URLs:
|]  http://10.36.255.86:19899/jobhistory/attempts/job_1593660756006_472283/m/KILLED
|]-----
|]Diagnostic Messages for this Task:
|]AttemptID:attempt_1593660756006_472283_m_001102_3 Timed out after 600 secs
```

打开时报的连接：FAILED TASK URLS，点到JOB的overview界面查看，可以看到map中的任务失败了。



所以结合报错信息可以定位，是map阶段的任务内存爆掉了导致任务失败。

先调大map任务的内存试试 set mapreduce.map.memory.mb=6144。结果正常运行。

虽然程序运行正常了，但是为什么20200701的任务在正常运行，到20200702的任务就map内存就爆掉了呢?

可以通过上面介绍的hive.explain.string的方法找到该任务的执行计划，通过对该任务的执行计划定位到问题SQL。

```sql
-- 播放及后续动作指标
select  user_id
        ,product
        ,0                                              as show_cnt
        ,0                                              as show_photo_num
        ,sum(play_cnt)                                  as play_cnt
        ,count(if(play_cnt > 0,1,null))                 as play_photo_num
        ,sum(play_duration)                             as play_duration
        ,sum(complete_play_cnt)                         as complete_play_cnt
        ,count(if(complete_play_cnt > 0,1,null))        as complete_play_photo_num
        ,sum(valid_play_cnt)                            as valid_play_cnt
        ,avg(play_progress)                             as play_progress
        ,sum(comment_stay_duration)                     as comment_stay_duration
        ,sum(like_cnt)                                  as like_cnt
        ,sum(if(like_cnt>0,1,null))                     as like_photo_num
        ,sum(click_like_cnt)                            as click_like_cnt
        ,sum(double_click_cnt)                          as double_click_cnt
        ,sum(cancel_like_cnt)                           as cancel_like_cnt
        ,sum(if(cancel_like_cnt>0,1,null))              as cancel_like_photo_num
        ,sum(comment_cnt)                               as comment_cnt
        ,sum(if(comment_cnt>0,1,null))                  as comment_photo_num
        ,sum(direct_comment_cnt)                        as direct_comment_cnt
        ,sum(reply_comment_cnt)                         as reply_comment_cnt
        ,sum(delete_comment_cnt)                        as delete_comment_cnt
        ,sum(if(delete_comment_cnt>0,1,null))           as delete_comment_photo_num
        ,0                                              as comment_like_cnt
        ,0                                              as comment_like_photo_num
        ,sum(follow_cnt)                                as follow_cnt
        ,sum(if(follow_cnt>0,1,null))                   as follow_photo_num
        ,sum(cancel_follow_cnt)                         as cancel_follow_cnt
        ,sum(if(cancel_follow_cnt>0,1,null))            as cancel_follow_photo_num
        ,sum(share_cnt)                                 as share_cnt
        ,sum(if(share_cnt>0,1,null))                    as share_photo_num
        ,0                                              as download_cnt
        ,0                                              as download_photo_num
        ,0                                              as report_cnt
        ,0                                              as report_photo_num
        ,0                                              as reduce_similar_cnt
        ,0                                              as reduce_similar_photo_num
from(
        select  nvl(user_id,0)                          as user_id
                ,p_product product
                ,photo_id
                ,sum(play_cnt)                          as play_cnt
                ,sum(play_duration)                     as play_duration
                ,sum(complete_play_cnt)                 as complete_play_cnt
                ,sum(valid_play_cnt)                    as valid_play_cnt
                ,avg(progress_rate)                     as play_progress
                ,sum(comment_stay_duration)             as comment_stay_duration
                ,sum(like_cnt)                          as like_cnt
                ,sum(click_like_cnt)                    as click_like_cnt
                ,sum(double_click_like_cnt)             as double_click_cnt
                ,sum(cancel_like_cnt)                   as cancel_like_cnt
                ,sum(comment_cnt)                       as comment_cnt
                ,sum(first_level_comment_cnt)           as direct_comment_cnt
                ,sum(second_level_comment_cnt)          as reply_comment_cnt
                ,sum(cancel_comment_cnt)                as delete_comment_cnt
                ,sum(follow_cnt)                        as follow_cnt
                ,sum(cancel_follow_cnt)                 as cancel_follow_cnt
                ,sum(share_success_cnt)                 as share_cnt
        from    ks_dw_aggr.party_ksprod_photo_consume_user_behv_di
        where   p_date= '{{ds_nodash}}'
        and     photo_id > 0
```

```
        and     photo_id > 0
        group by nvl(user_id,0)
                ,p_product
                ,photo_id
        )t2
```

可以看到问题SQL跟之前数据倾斜SQL基本一致，自然联想到问题会不会同样因为user_id=0 or user_id is null的热点数据导致。

然后我们对该表的20200701号数据以及20200702号的数据进行了对比。（20200701任务正常，20200702任务出错）。

date	nvl(user_id,0) = 0	total
20200702	1046797732	36576406139
20200701	972433558	35625278459
差值	74364174	951127680

可以看到20200702的数据比20200701的数据多了 9.5亿条，同时nvl(user_id,0) = 0的数据多了7400w条，因为数据量剧增导致的map内存爆掉。

在单独测试该任务时，运行缓慢，也同样出现了map阶段的数据倾斜问题。

```
INFO : 2020-07-03 17:19:38,175 Stage-1(job_1593660756006_697618) map = 100%, reduce = 94%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 736515.26 sec
INFO : 2020-07-03 17:19:41,222 Stage-1(job_1593660756006_697618) map = 100%, reduce = 95%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 736515.26 sec
INFO : 2020-07-03 17:19:47,375 Stage-1(job_1593660756006_697618) map = 100%, reduce = 96%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 738367.73 sec
INFO : 2020-07-03 17:19:53,547 Stage-1(job_1593660756006_697618) map = 100%, reduce = 97%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 739802.24 sec
INFO : 2020-07-03 17:20:02,790 Stage-1(job_1593660756006_697618) map = 100%, reduce = 98%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 741837.82 sec
INFO : 2020-07-03 17:20:24,298 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 743295.28 sec
INFO : 2020-07-03 17:21:25,834 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 745361.08 sec
INFO : 2020-07-03 17:22:27,233 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 746174.4 sec
INFO : 2020-07-03 17:23:28,774 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 746537.92 sec
INFO : 2020-07-03 17:24:30,179 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 746733.43 sec
INFO : 2020-07-03 17:25:31,660 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 746886.57 sec
INFO : 2020-07-03 17:26:33,098 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 747010.38 sec
INFO : 2020-07-03 17:27:34,495 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 747133.89 sec
INFO : 2020-07-03 17:28:35,900 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 747258.06 sec
INFO : 2020-07-03 17:29:37,447 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 747383.07 sec
INFO : 2020-07-03 17:30:38,851 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 747501.65 sec
INFO : 2020-07-03 17:31:40,190 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 747601.14 sec
INFO : 2020-07-03 17:32:41,598 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 747656.76 sec
INFO : 2020-07-03 17:33:43,131 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 747722.22 sec
INFO : 2020-07-03 17:34:44,719 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 747784.54 sec
INFO : 2020-07-03 17:35:46,114 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 747846.99 sec
INFO : 2020-07-03 17:36:47,609 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 747909.26 sec
INFO : 2020-07-03 17:37:49,155 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 747971.47 sec
INFO : 2020-07-03 17:38:50,502 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 748033.72 sec
INFO : 2020-07-03 17:39:51,832 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 748095.45 sec
INFO : 2020-07-03 17:40:53,252 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 748157.33 sec
INFO : 2020-07-03 17:41:54,640 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 748222.45 sec
INFO : 2020-07-03 17:42:56,027 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 748285.96 sec
INFO : 2020-07-03 17:43:57,427 Stage-1(job_1593660756006_697618) map = 100%, reduce = 99%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 748346.98 sec
INFO : 2020-07-03 17:44:22,072 Stage-1(job_1593660756006_697618) map = 100%, reduce = 100%; (pending maps/total maps : 0/1951, pending reduces/total reduces : 0/931), Cumulative CPU 748371.67 sec
```

经过讨论，同样采取业务切割的方式解决。