一、在会员分析中计算最近七天连续三天活跃会员数

最近七天连续三天活跃会员数 是一个描述性指标，应该在 ads 层建表：

```
drop table if exists ads.ads_member_continuous_active_count;
create table ads.ads_member_continuous_active_count
(
    `duration`  int comment '最近几天',
    `days`      int comment '活跃天数',
    `count`     bigint comment '活跃会员数'
) comment '持续活跃会员数'
partitioned by (`dt` string)
row format delimited fields terminated by ',';
```

并从 dws 层，会员每日启动汇总表 `dws.dws_member_start_day` 中统计：

```
with tmp1 as (
    -- 计算连续活跃的 start 日志，并设置 flag
    select device_id,
           dt,
           if(lag(dt) over (partition by device_id order by dt) is null,
               0,
               datediff(dt, lag(dt) over (partition by device_id order by dt)) - 1) as flag
    from dws.dws_member_start_day
    where dt <= '$do_date'
      and dt > date_add('$do_date', -7)
    distribute by device_id
    sort by dt
),
tmp2 as (
    -- 连续活跃的 start 设置相同的 session 标志
    select device_id,
           dt,
           sum(flag) over (partition by device_id order by dt) as session
    from tmp1
),
tmp3 as (
    select device_id,             -- 设备id
           min(dt)  as start_day, -- 持续活跃开始日期
           max(dt)  as end_day,   -- 持续活跃结束日期
           count(*) as days       -- 持续活跃天数
    from tmp2
    group by device_id, session
    having count(*) >= 3   -- 连续活跃3天及以上
)
--- 汇总并写入结果
insert overwrite table ads.ads_member_continuous_active_count
partition(dt='$do_date')
select 7 as duration,       -- 最近 7 日
```

```
       3 as days,           -- 连续 3 日
       count(distinct device_id) as count -- 活跃会员数
from tmp3;
```

统计结果：

duration	days	count	dt
7	3	2000	2020-07-23

二、项目的数据采集过程中，有哪些地方能够优化，如何实现？

Flume 单节点处理日志，且 channel 为 memory ，没有容错性且吞吐量不高，容易称为系统的瓶颈。改进方案：

1. 为 Flume 配置 LoadBlance 模式，提供系统容错性 Agent 节点的 flume_loadbalance_agent.conf 配置

```
agentX.sources = sX
agentX.channels = chX
agentX.sinks = sk1 sk2
agentX.sources.sX.channels = chX
agentX.sources.sX.type = exec
agentX.sources.sX.command = tail -F /data/logs/command.log
agentX.channels.chX.type = memory
agentX.channels.chX.capacity = 1000
agentX.channels.chX.transactionCapacity = 100
# Configure sinks
agentX.sinks.sk1.channel = chX
agentX.sinks.sk1.type = avro
agentX.sinks.sk1.hostname = 10.10.1.46
agentX.sinks.sk1.port = 44441
agentX.sinks.sk2.channel = chX
agentX.sinks.sk2.type = avro
agentX.sinks.sk2.hostname = 10.10.1.46
agentX.sinks.sk2.port = 44442
# Configure loadbalance
agentX.sinkgroups = g1
agentX.sinkgroups.g1.sinks = sk1 sk2
agentX.sinkgroups.g1.processor.type = load_balance
agentX.sinkgroups.g1.processor.backoff=true
agentX.sinkgroups.g1.processor.selector=round_robin
```

Collector1 节点的 flume_loadbalance_collector1.conf 配置

```
agent1.sources = s1
agent1.channels = ch1
agent1.sinks = sk1
agent1.sources.s1.channels = ch1
agent1.sources.s1.type = avro
agent1.sources.s1.bind = 10.10.1.46
agent1.sources.s1.port = 44441
agent1.channels.ch1.type = memory
agent1.channels.ch1.capacity = 1000
agent1.channels.ch1.transactionCapacity = 100
agent1.sinks.sk1.channel = ch1
agent1.sinks.sk1.type = logger
```

Collector2 节点的配置

```
agent2.sources = s2
agent2.channels = ch2
agent2.sinks = sk2
agent2.sources.s2.channels = ch2
agent2.sources.s2.type = avro
agent2.sources.s2.bind = 10.10.1.23
agent2.sources.s2.port = 44442
agent2.channels.ch2.type = memory
agent2.channels.ch2.capacity = 1000
agent2.channels.ch2.transactionCapacity = 100
agent2.sinks.sk2.channel = ch2
agent2.sinks.sk2.type = logger
```

启动 Flume

```
# 启动采集端，AgentX
$ ./bin/flume-ng agent --conf ./conf/ -f conf/flume_loadbalance_agent.conf -
Dflume.root.logger=DEBUG,console -n agentX
# 启动2个Collect端，Collector1和Collector2
$ ./bin/flume-ng agent --conf ./conf/ -f conf/flume_loadbalance_collector1.conf -
Dflume.root.logger=DEBUG,console -n agent1
$ ./bin/flume-ng agent --conf ./conf/ -f conf/flume_loadbalance_collector2.conf -
Dflume.root.logger=DEBUG,console -n agent2
```

1. 将 channel 修改为 kafka，利用 kafka 集群提高日志吞吐量，且当 flume 节点出现故障时，也不会丢失内存中未落盘的日志

```
a1.channels.channel1.type = org.apache.flume.channel.kafka.KafkaChannel
a1.channels.channel1.kafka.bootstrap.servers = kafka-1:9092,kafka-2:9092,kafka-3:9092
a1.channels.channel1.kafka.topic = channel1
a1.channels.channel1.kafka.consumer.group.id = flume-consumer
```