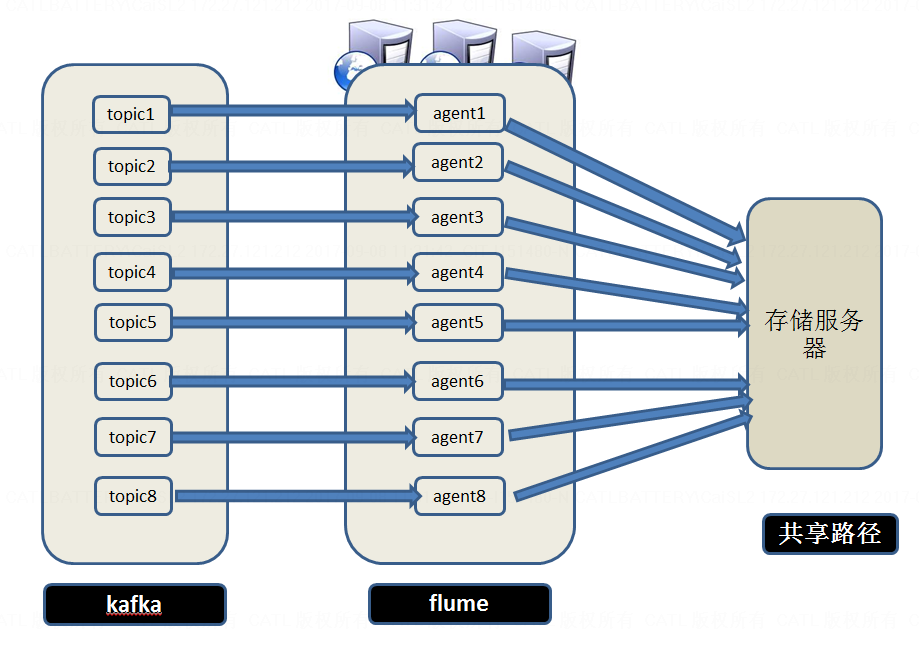
TES数据归档

1. 背景

TES测试系统测试数据目前全部存储在kafka,为了进一步加强对TES测试数据的管理和归档，经商议确认采用flume对kafka的数据进行以remark为单位按天归档，以.gz格式存储在磁盘中。

1. 架构图



1. Flume操作

Tes项目共有8个topic在生产消费数据,所以会起8个flume进程，每个进程针对一个topic的数据进行收集归档，每天凌晨将前一天的所有数据进行压缩打包，以下拿一个flume进程进行操作说明。

1. 本次采用flume归档，因最终存储采用压缩方式，所以自定义实现了flume 收集过程，将flume-service.jar加入到/usr/hdp/2.5.3.0-37/flume/lib目录下。
2. 编辑配置文件
3. vi /etc/flume/conf/flume-zip1.conf

#-------agent------

# define sources

agent.sources = kafkaSource

# define channels

agent.channels = memoryChannel

# define sinks

agent.sinks = sink

agent.sources.kafkaSource.channels = memoryChannel

agent.sinks.sink.channel = memoryChannel

#-------- kafkaSource-----------------

agent.sources.kafkaSource.type= org.apache.flume.source.kafka.KafkaSource

agent.sources.kafkaSource.zookeeperConnect= 172.26.38.74:2181,172.26.38.71:2181,172.26.38.72:2181

agent.sources.kafkaSource.topic = TES\_PRODUCT01

agent.sources.kafkaSource.kafka.consumer.timeout.ms = 1000

agent.sources.kafkaSource.interceptors = i

#agent.sources.kafkaSource.interceptors.i.type = timestamp

agent.sources.kafkaSource.interceptors.i.type=com.cn.hand.flume.ExtInterceptor$Builder

#------- memoryChannel-------------------------

agent.channels.memoryChannel.type = memory

agent.channels.memoryChannel.capacity=10000

agent.channels.memoryChannel.transactionCapacity=1000

#---------localSink ------------------

agent.sinks.sink.type = com.cn.hand.flume.SinkToFile

agent.sinks.sink.batchSize = 1000

#agent.sinks.sink.sink.serializer=TEXT

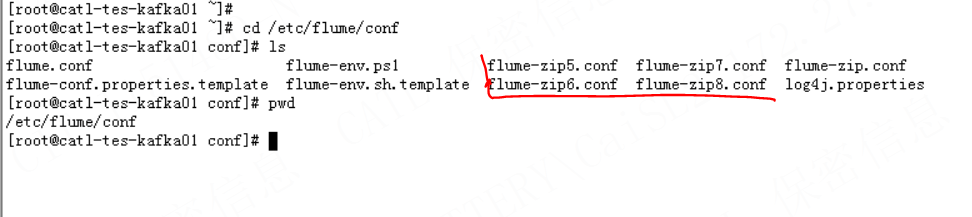
agent.sinks.sink.file.path = /tesdata

agent.sinks.sink.file.empPath = /flumedata/empdata

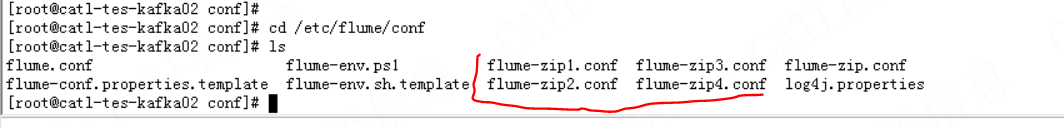
agent.sinks.sink.file.txnEventMax = 10000

agent.sinks.sink.file.maxOpenFiles = 5

机器1



机器2



1. flume启动后台命令

在72和71两台机器各写一个启动脚本,每个脚本启动4个flume进程。

72机器脚本编辑

**vi startup.sh**

flume-ng agent -c /etc/flume/conf -f /etc/flume/conf/flume-zip1.conf -Dflume.monitoring.port=19000 -Dflume.root.logger=DEBUG,console -n agent >/var/log/flumetes/flume1.log 2>&1 &

flume-ng agent -c /etc/flume/conf -f /etc/flume/conf/flume-zip2.conf -Dflume.monitoring.port=19001 -Dflume.root.logger=DEBUG,console -n agent >/var/log/flumetes/flume2.log 2>&1 &

flume-ng agent -c /etc/flume/conf -f /etc/flume/conf/flume-zip3.conf -Dflume.monitoring.port=19002 -Dflume.root.logger=DEBUG,console -n agent >/var/log/flumetes/flume3.log 2>&1 &

flume-ng agent -c /etc/flume/conf -f /etc/flume/conf/flume-zip4.conf -Dflume.monitoring.port=19003 -Dflume.root.logger=DEBUG,console -n agent >/var/log/flumetes/flume4.log 2>&1 &

71机器脚本编辑

**vi startup.sh**

flume-ng agent -c /etc/flume/conf -f /etc/flume/conf/flume-zip5.conf -Dflume.monitoring.port=19004 -Dflume.root.logger=DEBUG,console -n agent >/var/log/flumetes/flume5.log 2>&1 &

flume-ng agent -c /etc/flume/conf -f /etc/flume/conf/flume-zip6.conf -Dflume.monitoring.port=19005 -Dflume.root.logger=DEBUG,console -n agent >/var/log/flumetes/flume6.log 2>&1 &

flume-ng agent -c /etc/flume/conf -f /etc/flume/conf/flume-zip7.conf -Dflume.monitoring.port=19006 -Dflume.root.logger=DEBUG,console -n agent >/var/log/flumetes/flume7.log 2>&1 &

flume-ng agent -c /etc/flume/conf -f /etc/flume/conf/flume-zip8.conf -Dflume.monitoring.port=19007 -Dflume.root.logger=DEBUG,console -n agent >/var/log/flumetes/flume8.log 2>&1 &

启动完成后,查看监控端口是否存在

ps –ef | grep 进程号

1. 注:在同一台服务器启动多个flume进程，jar只需添加一次，一个flume进程实例对应一个配置文件和一个topic，即需要8个配置文件，以上红色字体内容需要相应修改。
2. 查看每个进程对应的日志

tail -200f /var/log/tesflume/flume1.log1

1. 端口规范：为便于8个flume进程的查看管理，启动时候port范围为19000~19008
2. Flume进程停止脚本

**vi stop.sh**

#!/bin/sh

#kill process

if [ $# -lt 1 ]

then

echo "no produrce\_name"

exit 1

fi

PROCESS=`ps -ef|grep $1|grep -v grep|grep -v PPID|awk '{ print $2}'`

for i in $PROCESS

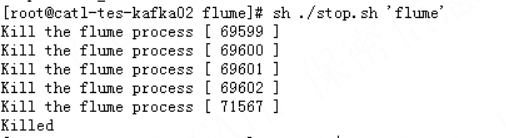
do

echo "Kill the $1 process [ $i ]"

kill -9 $i

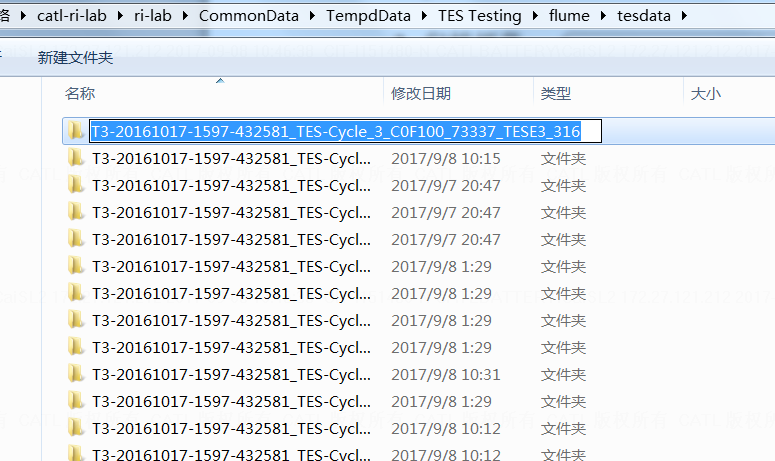
done

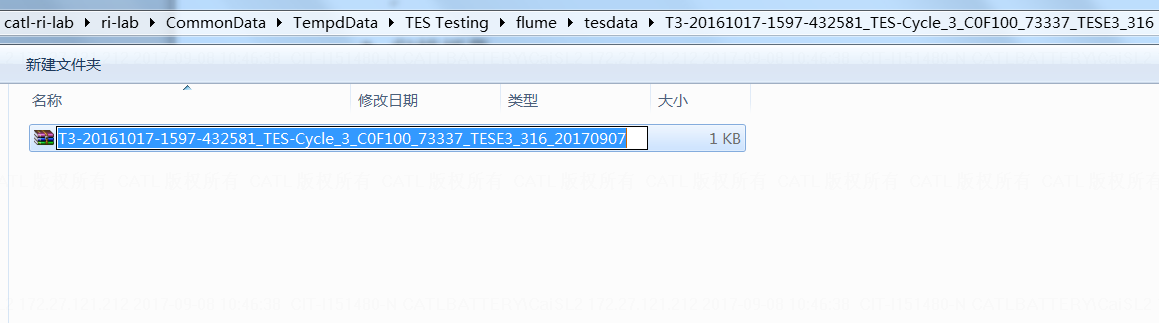
启动:sh ./stop.sh ‘flume’



1. 归档规范
2. 存储路径:存储路径由根/tesdata/remark/组成。
3. 压缩文件命名:每个压缩文件对应的一个remark一天的测试数据，以remark+当前存储日期命名。

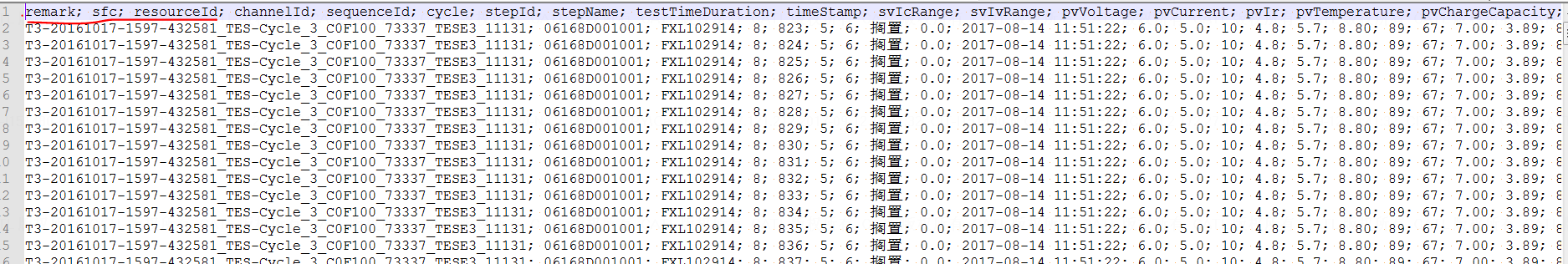
Eg:T3-20161017-1597-43281\_TES-Cycle\_3\_C0F100\_73337\_TESE3\_18\_20170907.gz





1. 解压后文档格式

解压后文档第一行为所有字段,接下来每一行为每一条测试数据对应字段的值,以分号隔开,以便于业务部导入excel查看。



1. 删除中间目录的空文件夹
2. 编写脚本

**vi clean\_dir.py**

# -\*- coding: utf-8 -\*-

import os

path = "/storage/flumedata/empdata"

files = os.listdir(path)

for f in files:

subpath = str(os.path.join(path, f)).strip()

if (os.path.isdir(subpath) and len(subpath)>10):

if (f[0] != '.' and os.path.getsize(subpath) == 4096):

try:

#print subpath

os.rmdir(subpath)

except BaseException:

print "delete error"

1. 定时任务启动脚本,每隔10天启动脚本清空目录

