**Flume数据采集：**

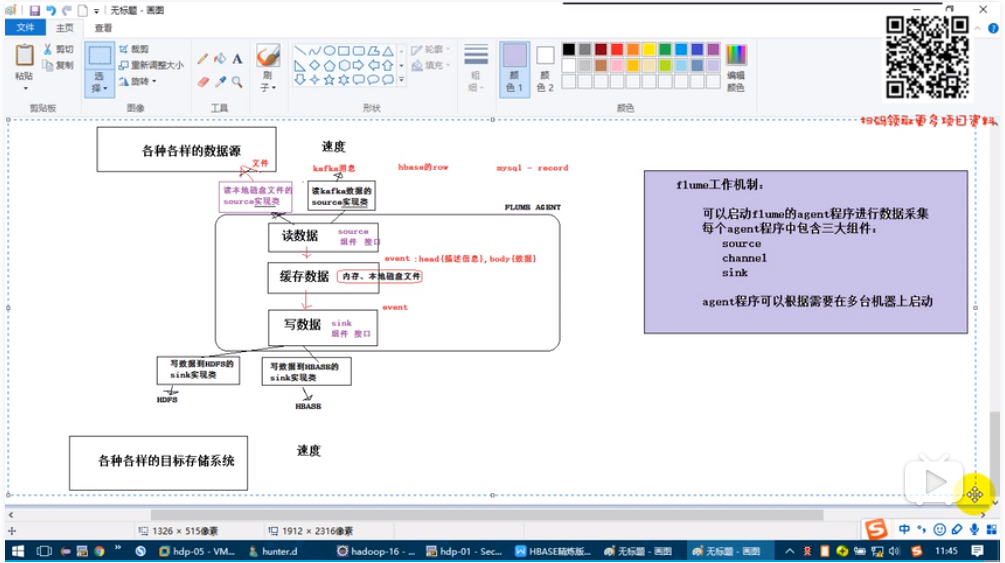
**原理：从各种各样数据源（业务系统）读取数据，因为是各种各样的数据源，包括文件系统，hdfs文件，hbase数据row，mysql 记录record，kafka消息等等，flume有一个组件叫source，用来读数据，读数据时，会把数据源转换成一个Event事件，Event事件是一个数据流单元，再用channel缓存数据到本地（内存，本地磁盘文件），为什么要缓存到本地呢，因为要保证写数据的恒定，因为数据源就是业务系统，可能某个时间段，业务系统大量被访问，产生大量日志，缓存到本地可以保证写数据的恒定；写数据用的是flume的sink组件，sink组件会将Event事件解析成存储系统的数据，这一套下来是flume的采集程序agent，数据较大时，可以配置多个agent一起来采集。所以flume数据采集。不用写代码，而是写配置文件！**

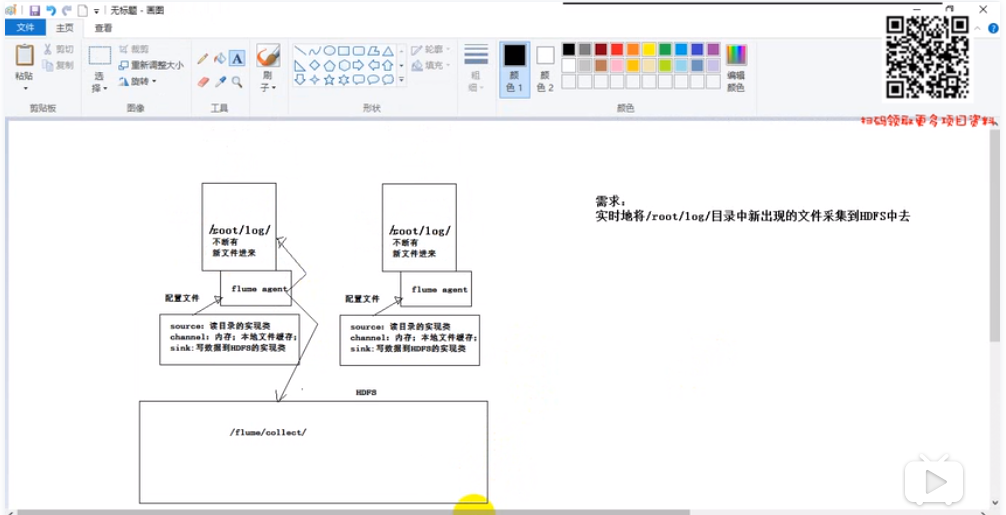
**Flume工作机制：**

**可以启动flume的agent程序进行数据采集**

**每个agent程序包含三大组件：source，channel，sink**

**Agent程序可以根据需要在很多台机器上启动**





Agent配置：

首先定义三大组件名称：

Ag1.sources = source1（ag1是agent）

Ag1.sinks= sink1

Ag1.channels = channel1

配置source组件(定义各种参数)

Ag1.sources.source1

配置sink组件(定义各种参数)

Ag1.sinks.sink1

配置channel组件(定义各种参数)

Ag1.channels.channel1

**启动: bin/flume-ng agent**

**-c conf/ 是指flume的配置文件conf，**

**-f dir-hdfs.conf 是采集配置的文件**

**-n a1 是指自己定义的agent名字**

**-Dflume.root.logger=DEBUG,console 是将日志打印到控制台（不然会将日志打印到文件，不便于观察）**

1. **后台启动：nohup bin/fl dir-hdfs（从本地文件读数据，写入hdfs文件）**

**# spooldir source监控目录，采集的是静态文件**

a1.sources = source1

a1.sinks = sink1

a1.channels = channel1

# 配置source组件

a1.sources.source1.type = spooldir #读取数据的类型，spooldir表示本地磁盘上的文件

a1.sources.source1.spoolDir= /opt/module/logs #本地磁盘上的文件位置

a1.sources.source1.filesuffix = .FINISHEO

a1.sources.source1.deserializer.maxLin eLength = 5120

# 配置sink组件

a1.sinks.sink1.type = hdfs #写数据的类型

a1.sinks.sink1.hdfs.path= hdfs://h1:9000/dj/access\_log/%y-%m-%d/%H-%M #写数据的hdfs文件位置

a1.sinks.sink1.hdfs.filePrefix= dj\_log #生成文件的前缀

a1.sinks.sink1.hdfs.fileSuffix= .log #生成文件的后缀

a1.sinks.sink1.hdfs.batchSize = 100 #读多少条event后写hdfs

a1.sinks.sink1.hdfs.fileType= DataStream #读的什么文件，写就什么文件

a1.sinks.sink1.hdfs.writeFormat = Text

#roll:滚动切换：控制写文件的切换规则(设0就是不按这个规则切换)

a1.sinks.sink1.hdfs.rollSize = 102400 #按文件字节

a1.sinks.sink1.hdfs.rollCount = 1000000 #按event条数切

a1.sinks.sink1.hdfs.rollInterval = 60 #按时间间隔切换文件

#控制生成目录的规则（1分钟）

a1.sinks.sink1.hdfs.round = true

a1.sinks.sink1.hdfs.roundValue = 1

a1.sinks.sink1.hdfs.roundUnit = minute

a1.sinks.sink1.hdfs.useLocalTimeStamp = true #获取当前机器的时间

#配置channel组件

a1.channels.channel1.type = memory

a1.channels.channel1.capacity = 500000 #event条数

a1.channels.channel1.transactionCapacity = 600 #flume事务控制所需要的缓存容量600条event

#绑定source，channel和sink之间的连接

a1.sources.source1.channels = channel1

a1.sinks.sink1.channel = channel1

1. **exec\_file\_hdfs.conf**

**# exec source 可以实时搜集数据**

#定义各个组件

agent1.sources = src

agent1.channels = ch

agent1.sinks = des

#配置source

agent1.sources.src.type = exec

agent1.sources.src.command = tail -F /data/flume2/goods

#配置channel

agent1.channels.ch.type = file

agent1.channels.ch.checkpointDir = /data/flume2/ckdir

agent1.channels.ch.dataDirs = /data/flume2/datadir

#配置sink

agent1.sinks.des.type = hdfs

agent1.sinks.des.hdfs.path = hdfs://localhost:9000/myflume2/exec\_file\_hdfs/%Y%m%d/

agent1.sinks.des.hdfs.useLocalTimeStamp = true

#设置flume临时文件的前缀为 . 或 \_ 在hive加载时，会忽略此文件。

agent1.sinks.des.hdfs.inUsePrefix=\_

#设置flume写入文件的前缀是什么

agent1.sinks.des.hdfs.filePrefix = abc

agent1.sinks.des.hdfs.fileType = DataStream

agent1.sinks.des.hdfs.writeFormat = Text

#hdfs创建多久会新建一个文件，0为不基于时间判断,单位为秒

agent1.sinks.des.hdfs.rollInterval = 30

#hdfs

写入的文件达到多大时，创建新文件 0为不基于空间大小,单位B

agent1.sinks.des.hdfs.rollSize = 1

#hdfs有多少条消息记录时，创建文件，0为不基于条数判断

agent1.sinks.des.hdfs.rollCount = 1

#hdfs空闲多久就新建一个文件,单位秒

agent1.sinks.des.hdfs.idleTimeout = 60

##下面是把上面设置的组件关联起来（把点用线连起来）

agent1.sources.src.channels = ch

agent1.sinks.des.channel = ch

1. **exec\_mem\_hdfs.conf**

#定义各个组件

agent1.sources = src

agent1.channels = ch

agent1.sinks = des

#配置source

agent1.sources.src.type = exec

agent1.sources.src.command = tail -n 20 /data/flume2/goods

#配置channel

agent1.channels.ch.type = memory

agent1.channels.ch.keep-alive = 30

agnet1.channels.ch.capacity = 1000000

agent1.channels.ch.transactionCapacity = 100

#配置sink

agent1.sinks.des.type = hdfs

agent1.sinks.des.hdfs.path = hdfs://localhost:9000/myflume2/exec\_mem\_hdfs/%Y%m%d/

agent1.sinks.des.hdfs.useLocalTimeStamp = true

#设置flume临时文件的前缀为 . 或 \_ 在hive加载时，会忽略此文件。

agent1.sinks.des.hdfs.inUsePrefix=\_

#设置flume写入文件的前缀是什么

agent1.sinks.des.hdfs.filePrefix = abc

agent1.sinks.des.hdfs.fileType = DataStream

agent1.sinks.des.hdfs.writeFormat = Text

#hdfs创建多久会新建一个文件，0为不基于时间判断,单位为秒

agent1.sinks.des.hdfs.rollInterval = 30

#hdfs写入的文件达到多大时，创建新文件 0为不基于空间大小,单位B

agent1.sinks.des.hdfs.rollSize = 1

#hdfs有多少条消息记录时，创建文件，0为不基于条数判断

agent1.sinks.des.hdfs.rollCount = 1

#hdfs空闲多久就新建一个文件,单位秒

agent1.sinks.des.hdfs.idleTimeout = 30

##下面是把上面设置的组件关联起来（把点用线连起来）

agent1.sources.src.channels = ch

agent1.sinks.des.channel = ch

1. **exec\_mem\_logger.conf**

#定义各个组件

agent1.sources = src

agent1.channels = ch

agent1.sinks = des

#配置source

agent1.sources.src.type = exec

agent1.sources.src.command = tail -n 20 /data/mydata/goods

#配置channel

agent1.channels.ch.type = memory

#配置sink

agent1.sinks.des.type = logger

##下面是把上面设置的组件关联起来（把点用线连起来）

agent1.sources.src.channels = ch

agent1.sinks.des.channel = ch

1. **syslog\_mem\_logger.conf**

#定义各个组件

agent1.sources = src

agent1.channels = ch

agent1.sinks = des

#配置source

agent1.sources.src.type = syslogtcp

agent1.sources.src.port = 6868

agent1.sources.src.host = localhost

#配置channel

agent1.channels.ch.type = memory

#配置sink

agent1.sinks.des.type = logger

##下面是把上面设置的组件关联起来（把点用线连起来）

agent1.sources.src.channels = ch

agent1.sinks.des.channel = ch

1. Source=avro 是多级agent串联。

7.

2.sources=s1

a2.channels=c1

a2.sinks.k1

a2.sources.s1.type=exec

a2.sources.s1.command=head -n 2 /usr/local/src/datas/a.txt

a2.channels.c1.type=memory

a2.sinks.k1.type=hive

a2.sinks.k1.hive.metastore=thrift://master:9083

a2.sinks.k1.hive.database=dj

a2.sinks.k1.hive.table=text

#DELIMITED：处理简单的分隔文本事件。

a2.sinks.k1.serializer=DELIMITED

#传入数据中的字段分隔符。要使用特殊字符，请用双引号将其括起来，如“\t”

a2.sinks.k1.serializer.delimiter="\t"

a2.sinks.k1.serializer.serdeSeparator='\t'

a2.sinks.k1.serializer.fieldnames=user\_id,user\_name,age

a2.sources.s1.channels=c1

a2.sinks.k1.channel=c1