Flume1.8总结

Flume是一个日志采集工具，它具有分布式、高可靠、高可用的特性，能够有效的收集、聚合、移动大量日志数据。高可靠是指flume将数据发送到下一个agent或存储平台且数据被接受后，存放在channel的数据才会被移除，flume使用事务进行数据的传输。

Flume的三个组件及其作用：

Source：从外界采集数据，可以处理各种类型、各种格式的日志数据，包括avro，thrift，exec，jms，spooling directory，netcat，kafka等。

Channel 对采集到的数据进行缓存，可以存放在Memory或File中，常见存储类型还有kafka channel，jdbc channel

Sink组件用于把数据发送到目的地，包括hdfs，logger，avro，thrift，ipc，file，hbase，自定义等。

我公司采用的Source类型为：监控后台日志：exec。

Tail -F 和tail -f区别

Tail -F 根据文件名进行追踪，并保持重试，当文件名被改名或删除后，如果再次创建相同的文件名，会继续追踪。

Tail -f 根据文件描述符进行追踪，当文件改名或被删除，追踪停止

Flume中组件之间传输的基本单位是Event。

Event的数据结构：包括header和body

Header是一个map结构，存储的是一个key-value类型

Body是一个字节数组，里面存储具体内容。

Flume之拦截器interceptors：拦截source发出的数据：对于一个source可以指定一个或多个拦截器，按先后顺序依次对数据进行处理。

常见的interceptors类型：timestamp interceptor，host interceptor，static interceptor，UUID interceptor，regex Filtering interceptor，regex extractor interceptor （正则抽取拦截器）

Search and Replace interceptor(查询和替换拦截器)。

自定义拦截器：实现interceptor接口，重写方法实现逻辑，配置自己的拦截器。

Channel：

使用memory channel性能比较高，但是数据不安全（jvm宕机，内存数据会丢失）

使用file channel数据安全，相对memoey channel来说性能较低。

Channel Selectors channel选择器：

可以让不同的项目日志通过不同channel到不同的sink中去。

官方的channel selectors有两种：复制选择器（Replicating Channel Selector默认），按照规则发送数据的拦截器：Multiplexing Channel Selector

Sink：数据最终目的地

常见sink类型：logger，hdfs，kafka，avro等

Hdfs sink需hdfs的配置文件和类库，一般采取多个sink汇聚到一台采集机器推送到hdfs（该flume节点需要安装hadoop-client）

Sink处理器 Processor 默认单个sink的时候Default sink Processor

一个agent中有多个sink的时候：故障转移选择器failover sink processor。

负载均衡选择器：load balancing sink processor （随机算法和轮询算法）

自定义组件：

自定义sink：继承abstractsink实现configurable，重写start和stop方法实现逻辑。

自定义source：

Source有两种PollableSource、EventDrivernSource。

PollableSource 有process方法处理，系统自动调用process（）

EventDrivernSource 没有process（），通过start（）来执行向channel中发送数据的操作。

步骤：configure读取用户配置信息，如果有process获取组装event，重写start和stop方法

Flume进程监控

由于flume是一个单进程程序，会存在单点故障，所以需要一个监控机制，发现flume进程dowm掉之后，需要重启。 通过shell脚本实现flume进程监控及自动重启。

Flume配置调整：

调整flume-env.sh脚本中的JAVA\_OPTS参数，设置JVM虚拟机的内存大小，建议设置1G-2G,太小的话会导致flume进程gc频繁。

Top

Jstat -gutil pid 1000 查看java进程的gc情况

在一台服务器上启动多个agent的时候，建议拷贝多个conf目录，修改log4j.properties中日志的存储目录（可以保证多个agent的日志分别存储），并且把日志级别调整为warn（减少垃圾日志的产生）

Flume 参数调优

1 Source 增加Source个数参数可以增大Source的读取数据的能力，例如：当某一个目录产生的文件过多时需要将这个文件目录拆分成多个文件目录，同时配置好多个Source，以保证Source有足够的能力获取新产生的数据。BatchSize参数决定source一次批量运输到channel的event条数，适当调大batchsize可以提高source搬运event到channel的性能。

2 channel type选择memory时 channel性能最好，但是如果flume进程意外挂掉可能会丢失数据。Type选择file时channel的容错性（或安全性）但是性能上会比memory channel差。使用file channel时，dataDirs配置多个不同盘下的目录可以提高性能。

Capacity参数决定channel可容纳最大event条数。TransactionCapacity参数决定每操source往channel里面写的最大event条数和每次sink从channel里面读的最大event条数。Transaction需要大于source和sink的batchsize参数。

3 sink增加sink的个数可以增加sink消费event的能力。Sink也不是越多越好，够用就行，过多的sink会占用系统资源，造成系统不必要的浪费。Batchsize参数决定sink一次批量从channel读取的event条数。适当调大这个参数可以提高sinl从channel搬出event的性能。

5 flume的事务机制

Flume使用两个独立的事务分别负责从source到channel（put）以及从channel到sink的事务传递（take）。

6 Flume采集数据会丢失吗？

不会 channel存储可以存储在File中，数据传输自身有事务。