# flume学习

## flume简介

是分布式、可靠、可用性好的服务，用于收集、聚合、移动大量日志数据。基于流计算的简单灵活架构。用于在线分析

### 优点

可以和任意集中式存储进程集成

输入的数据速率大于写入存储目的地的速率flume会进行缓冲

flume提供上下文路由(数据流路线)

flume中的事务基于channe,使用了两个事务模型(sender+receive)

### 特点

flume高效收集web server的log到hdfs

可以高效获取输入

导入大量数据

flume支持大量的source和destination类型

flume支持多级跳跃,source和destination的扇入(fan in)和扇出(fan out)

flume可以水平伸缩

### hdfs put的问题

同一时刻只能传输一个文件

put处理的是静态文件(不能实时被写入)

### flume架构

#### 描述

在数据生成器运行的节点上启动单独的flume agent,来收集数据.数据收集器收集数据,推送到hdfs

#### flume event

事件是flume的传输单元,主要是byte[],可以含有一些header信息,在source和destination之间

#### fluem agent

每个代理是一个独立的java进程(jvm),从客户端或者其他agent接收数据然后转发到下一个destination(sink|agent)

包含三个组件

##### source

从事件生成器接收数据, 以event的方式传给一个或多个channel,

##### channel

通道,从source中接收flume event,作为临时存放地(buffer,缓冲区),直到sink将其消费掉,是source和sink之间的桥梁

channel是事务的,可以和多个source或sink协同

##### sink

沉槽:存放数据到hdfs等中,从channel中消费event,并分发给destination.sink的destination可以是另一个sink也可以是一个hdfs(kafka,hbase,avro)等

##### 注意:一个flume的agent可以有多个source,channel,sink

#### flume的附加组件

Interceptors

Channel Selectors

Sink Processors

## 安装flume

### 下载apache-flume-1.6.0-bin.tar然后解压

### 修改文件

将conf下的flume-conf.properties.template重命名为flume-conf.properties,将flume-env.sh.template重命名为flume-env.sh

,修改flume-env.sh里面的java环境变量为java\_home的目录

### 配置flume

命名agent组件

agent可以有多个sources,多个channels,多个sinks

代理名称.sources=r1,r2(俩source中间逗号隔开)

代理名称.channels= c1,c2(俩channel中间逗号隔开)

代理名称.sinks= s1,s2(俩sink中间逗号隔开)

source-r1

代理名称.sources.r1.type=

代理名称.sources.r1.xxx=

代理名称.sources.r1.yyy=

sink-s1

代理名称.sinks.s1.type=

代理名称. sinks.s1.xxx=

channels -c1

代理名称.channels.c1.type=

代理名称. channels.c1.xxx=

binding,source可以配置多个channel

代理名称.sources.r1.channels=c1

sink只能配置一个channel

代理名称.sinks.s1.channel=c1

#### 简单实例

*# example.conf: A single-node Flume configuration*

*# Name the components on this agent*

a1.sources **=** r1

a1.sinks **=** k1

a1.channels **=** c1

*# Describe/configure the source*

a1.sources.r1.type **=** netcat

a1.sources.r1.bind **=** localhost

a1.sources.r1.port **=** 44444

*# Describe the sink*

a1.sinks.k1.type **=** logger

*# Use a channel which buffers events in memory*

a1.channels.c1.type **=** memory

a1.channels.c1.capacity **=** 1000

a1.channels.c1.transactionCapacity **=** 100

*# Bind the source and sink to the channel*

a1.sources.r1.channels **=** c1

a1.sinks.k1.channel **=** c1

#### 配置conf/ flume-conf.properties

#### 运行flume agent

flume-ng agent --conf conf --conf-file example.conf --name a1 -Dflume.root.logger=INFO,console

--conf :配置目录

--conf-file:配置文件

--name:代理名称

-D:指定额外的参数

flume-ng agent --conf /home/bigData/flume/flume/conf --conf-file flume-conf.properties --name a1 -Dflume.root.logger=INFO,console

## 核心类检查

### source

生成event,调用channelprocessor的方法,将event put到channel中

source🡪event🡪channelprocessor🡪Interceptorchain(e)🡪channels(e)

### sink

连接到Channel,消费里面的event,将其发送到destination,有很多相应的sink类型,sink可以根据sinkgroup和sinkprocessor进行分组,通过processor有sinkrunner轮询出来,sink的process()方法只能有一个线程访问

### channel

连接Source(Event Procuder)和sink(Event Consumer),本质上Channel就是Buffer.支持事务处理,保证原子性(put+take),channel是线程安全的

## http source

走http协议,默认使用json格式处理器