

Flume应用开发

www.huawei.com



目标

- 学完本课程后，您将能够：
 - 了解**Flume**应用开发适用场景
 - 掌握**Flume**应用开发



目录

1. Flume应用场景
2. Flume应用开发
3. Flume应用开发举例

Flume应用场景

- **Flume** 的核心是把数据从数据源收集过来，再送到目的地。为了保证输送一定成功，在送到目的地之前，会先缓存数据，待数据真正到达目的地后，删除自己缓存的数据。
- **Flume** 采用流式方式采集和传送数据，程序配置好后，不需要外部条件触发下，一直监控数据源，源源不断地采集、传送数据到目的地。
- 主要应用于以下几种场景：
 - 将分布式节点上大量数据实时采集、汇总和转移。
 - 将集群内、外的本地文件、实时数据流采集到**FusionInsight**集群内的**HDFS**、**Hbase**、**Kafka**、**Solr**中。
 - 将**Avro**、**Syslog**、**http**、**Thrift**、**JMS**、**Log4j**协议发送过来的数据采集到**FusionInsight**集群内。



目录

1. **Flume**应用场景
2. **Flume**应用开发
3. **Flume**应用开发举例

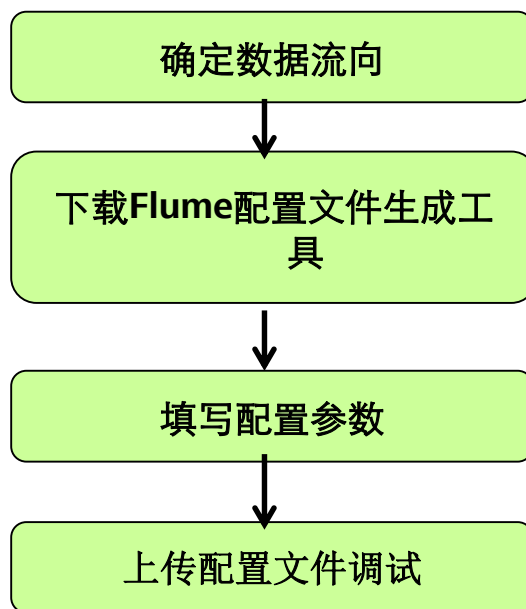
基本概念

- **Flume**基本概念：

- **Source**:数据源，即是产生日志信息的源头，**Flume**会将原始数据建模抽象成自己处理的数据对象：**event**。
- **Channel**:通道，主要作用是临时缓存**Source**发过来数据。
- **Sink**:主要作用是从**channel**中取出数据并将数据放到不同的目的地。
- **event**：一个数据单元，带有一个可选的消息头，**Flume** 传输的数据的基本单位是 **event**，如果是文本文件，通常是一行记录。**event** 从 **Source**，流向 **Channel**，再到 **Sink**，**Sink**将数据写入目的地。

流程简介

- **Flume**当前已经提供了采集数据的客户端，该客户端已经能满足用户数据采集场景，使用**Flume**的核心是需要根据采集数据场景，开发**Flume**的配置文件（客户端和服务端都需要该配置文件），开发流程如下：



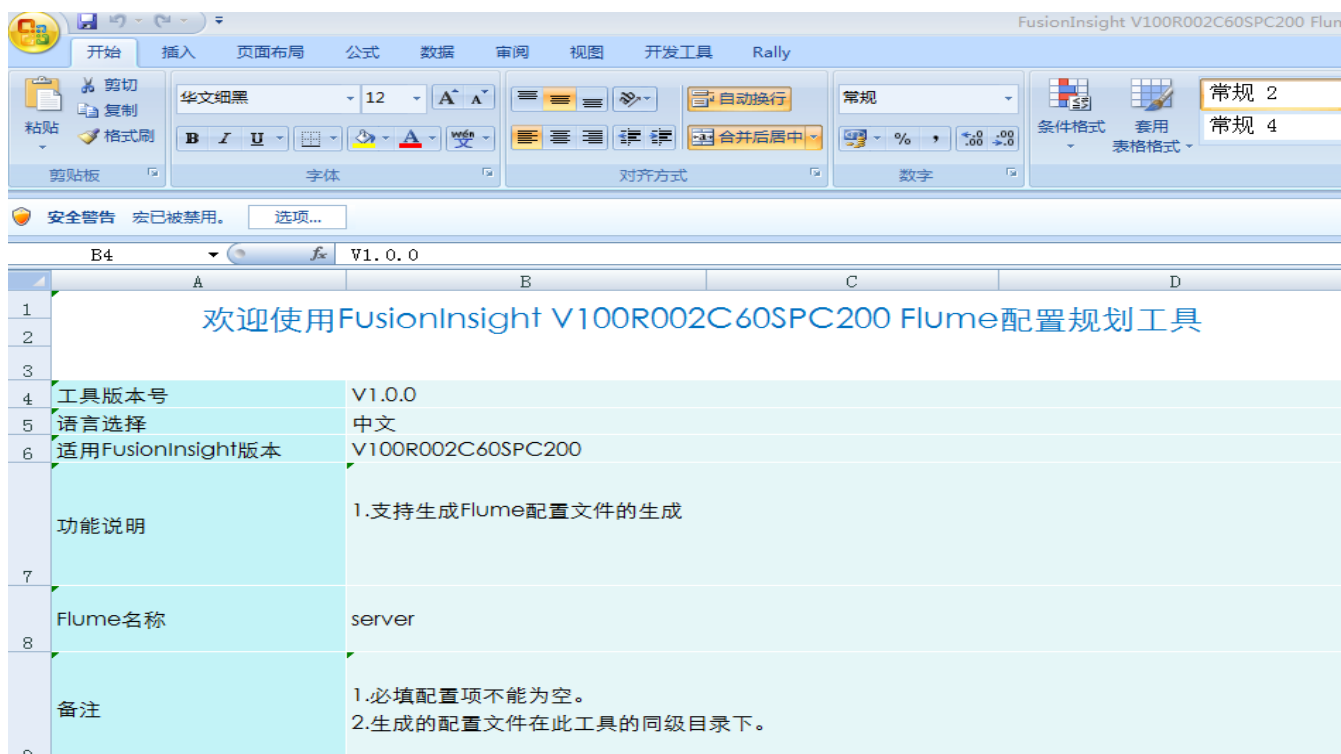
数据流向

确定数据流向

- 确定数据源是在集群内还是集群外，如果是集群内，那么可以直接通过**Flume**服务端采集，如果集群外，需要通过客户端将数据采集，然后通过级联的方式将数据发给**Flume**服务端。
- 确定数据最终去向：**HDFS**、**HBase**、**Kafka**、**Solr**。
- 根据采集源类型和数据最终流向确定**source/channle/sink**类型。

配置工具

从版本发布路径（中文：**ProductDoc\zh\pdf\FusionInsight V100R002C60SPC200 Flume配置规划工具 01.zip**）解压配置工具，取出**Flume**配置文件生成工具：**FusionInsight V100R002C60SPC200 Flume配置工具.xlsm**，打开后，工具如图：



	A	B	C	D
1	欢迎使用FusionInsight V100R002C60SPC200 Flume配置规划工具			
2				
3				
4	工具版本号	V1.0.0		
5	语言选择	中文		
6	适用FusionInsight版本	V100R002C60SPC200		
7	功能说明	1.支持生成Flume配置文件的生成		
8	Flume名称	server		
9	备注	1.必填配置项不能为空。 2.生成的配置文件在此工具的同级目录下。		

配置Flume名称

从填写模板中的“**Flume名称**”项，取值为：**server** 和**client**（分别表示生成客户端和服务端配置文件），如图：

工具版本号	V1.0.0
语言选择	中文
适用FusionInsight版本	V100R002C60SPC200
功能说明	1.支持生成Flume配置文件的生成
Flume名称	server

配置关键步骤

在“Flume配置” sheet页的主界面中，有“添加source”、“添加channel”、“添加sink”、“生成配置文件”按钮，点击上述3个按钮可以分别生成对应的配置项

A	B	C	D	E	F
添加Source			添加Channel		
Source配置项	Source配置描述	Source配置内容	Channel配置项	Channel配置描述	Channel配置内容

G	H	I	J	K	L
添加Sink			生成配置文件		
Sink配置项	Sink配置描述	Sink配置内容			

配置source (1)

添加source配置项：

点击界面中“添加source”按钮，先填写当前source名字，然后选择type配置项的下拉框，选择source的类型，如图

	A	B	C	D
1	添加Source			
2	Source配置项	Source配置描述	Source配置内容	Channel配置项
3	SourceName	Source名称，必须唯一	test1	
4	type	Source类型，取值为	spooldir	
5	spoolDir	待采集的文件所在的目录路径，此参数不能为	spooldir kafka http taildir avro	
		采集完成后的文件添加		

目前支持的类型有:spooldir（从某个目录下采集数据）、kafka（从kafka中采集数据）、http（接受http请求的数据）、taildir（实时采集目录下的文件）、avro（接受avro协议的数据），请根据当前Flume采集数据源情况选择type类型。

配置source (2)

添加source配置项：

当选择source某一个type类型后，工具会自动显示出该类型的配置项、配置项说明，请根据当前环境情况和配置规则填写，如图：

A	B	C
添加Source		
Source配置项	Source配置描述	Source配置内容
SourceName	Source名称，不能为空，必须唯一。	client_test
type	Source类型，取值为spooldir, kafka, http, taildir, avro中的任意一个	kafka
kafka.topics	订阅的kafka topic列表，用逗号分隔，此参数不能为空	
kafka.topics.regex	kafka topic正则表达式，符合正则表达式的topic会被订阅，优先级高于“kafka.topics”，如果存在将覆盖“kafka.topics”	
kafka.consumer.group.id	从kafka中获取数据的组标识，此参数不能为空	test

图中当选择type为kafka后，表示Flume从Kafka中采集数据，Kafka相关配置项便会显示出来，每一个配置项作用和配置规则都在配置描述中有相关说明，特别需要关注参数项不能为空的情况，例如左图中必填参数：

SourceName:是该source的名字，例如为:client_test。

type: 该source的类型，为kafka，表明该source是从Kafka中读取数据。

kafka.topics: 从Kafka哪些topic读取数据

kafka.consumer.group.id: Kafka读取数据的组ID标识，例如test。

其他的非必填参数请参照配置项说明填写。

配置channel

添加channel配置项：

点击界面中“添加channel”按钮，先填写当前channel名字，然后选择type配置项的下拉框，选择channel的类型，如图：

:

D	E	F	
添加Channel			
Channel配置项	Channel配置描述	Channel配置内容	Sim
ChanelName	Chanel名称，必须		
type	Chanel类型，取值	file	
dataDirs	缓冲区数据保存目录，默认为运行目录	file memory channel/data	
checkpointDir	checkpoint 信息保存目录，默认在运行目录下	~/.flume/file-channel/checkpoint	

目前支持的类型有:file、memory，上述配置项数据在传输过程中缓存在文件和内存，请根据数据可靠性要求选择其类型。

配置sink

添加sink配置项：

点击界面中“添加channel”按钮，先填写当前sink名字，然后选择type配置项的下拉框，选择sink的类型，如图：

添加Sink		
Sink配置项	Sink配置描述	Sink配置内容
SinkName	Sink名称，不能为空，必须唯一	client_sink
type	Sink类型，取值为	hdfs
hdfs.path	写入HDFS的目录，此参数不能为空	hdfs
hdfs.filePrefix	写入HDFS后文件的前缀	hbase
hdfs.fileSuffix	写入HDFS后文件的后缀	kafka
		avro
		solr

当前支持的类型有：

HDFS: 将数据写入HDFS

HBase:将数据写入HBase

Kafka:将数据写入Kafka

Avro:将数据发给下一跳的Flume

Solr:将数据写入Solr

请根据自己预先规划的数据流向目的地选择相应类型。

生成配置文件

在配置**source**、**channel**、**sink**后，请点击工具中“生成配置文件”按钮，会在当前模板工具下生成**properties.properties**文件

G	H	I	J	K	L	M	N
添加Sink			生成配置文件				
配置项	Sink配置描述	Sink配置内容					
名称	Sink名称，必须唯一						
	Sink类型，取值为	hbase					
	Hbase表名，此参数不能为空						
family	Hbase列族名，此参数不能为空						
batchSize	Flume一次写入Hbase中的最大事件数	100					
	是否批量设置Hbase列的						



目录

1. Flume应用场景
2. Flume应用开发
3. Flume应用开发举例

业务背景

举例：将集群外某一个节点上目录下的数据文件采集到**Kafka**中，集群已安装完毕，其他信息如下：

数据源：

节点：**192.168.1.20**

采集路径：**/tmp/flume_test**

服务端：**Flume**部署在 **192.168.1.200 192.168.1.201**

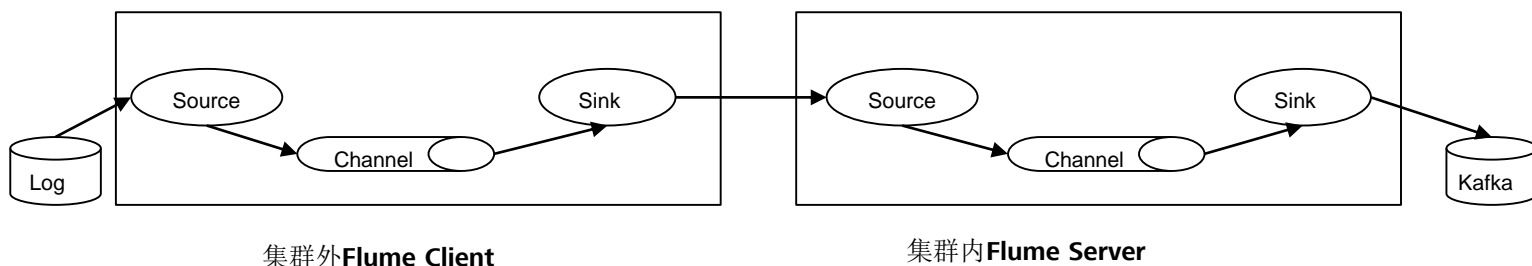
方案设计（1）

分析

1) 首先确定数据源：集群外

2) 数据最终流向：**Kafka**(集群内)

那么需要采用级联方式，使用**Flume** 客户端采集数据，**Flume**服务端接受数据，并将数据存储在**Kafka**中，数据流向如下图：



方案设计（2）

3) 根据采集源类型和数据最终流向确定**source/channel/sink**类型

客户端：

source: spool

channel:file (防止进程复位数据丢失)

sink:avro (级联**Flume** 服务端时使用)

服务端：

source:avro (级联**Flume** 客户端时使用)

channel:file

sink:Kafka (数据最终落地)

客户端配置—配置source

使用配置文件生成工具填写相应的配置参数，先生成客户端配置信息，其中**source**各项参数填写如下（其他参数值可以使用工具的默认值）：

SourceName	client_source
type	spooldir
spoolDir	/tmp/flume_test
trackerDir	/tmp/flume_tracker
channels	client_channel

客户端配置—配置channel

channel各项参数填写如下（其他参数值可以使用工具的默认值）：

ChanelName	client_channel
type	file
dataDirs	/tmp/file-channel/data
checkpointDir	/tmp/file-channel/checkpoint

客户端配置—配置sink

sink各项参数填写如下（其他参数值可以使用工具的默认值）：

SinkName	client_sink
type	avro
hostname	192.168.1.100
port	21154
channel	client_channel

点击“生成配置文件”按钮，将工具当前目录下的**properties.properties**配置文件上传到客户端安装目录下**/fusioninsight-flume-1.6.0/conf/**中。

服务端配置—配置source

服务端**source**修改的参数名和参数值为（其他参数值可以使用工具的默认值）：

SourceName	server_source
type	avro
bind	192.168.1.100
port	21154
channels	server_channel

服务端配置—配置channel

服务端**channel**修改的参数名和参数值为（其他参数值可以使用工具的默认值）：

ChanelName	server_channel
type	file
dataDirs	/tmp/server_file-channel/data
checkpointDir	/tmp/server_file-channel/checkpoint

服务端配置—配置sink

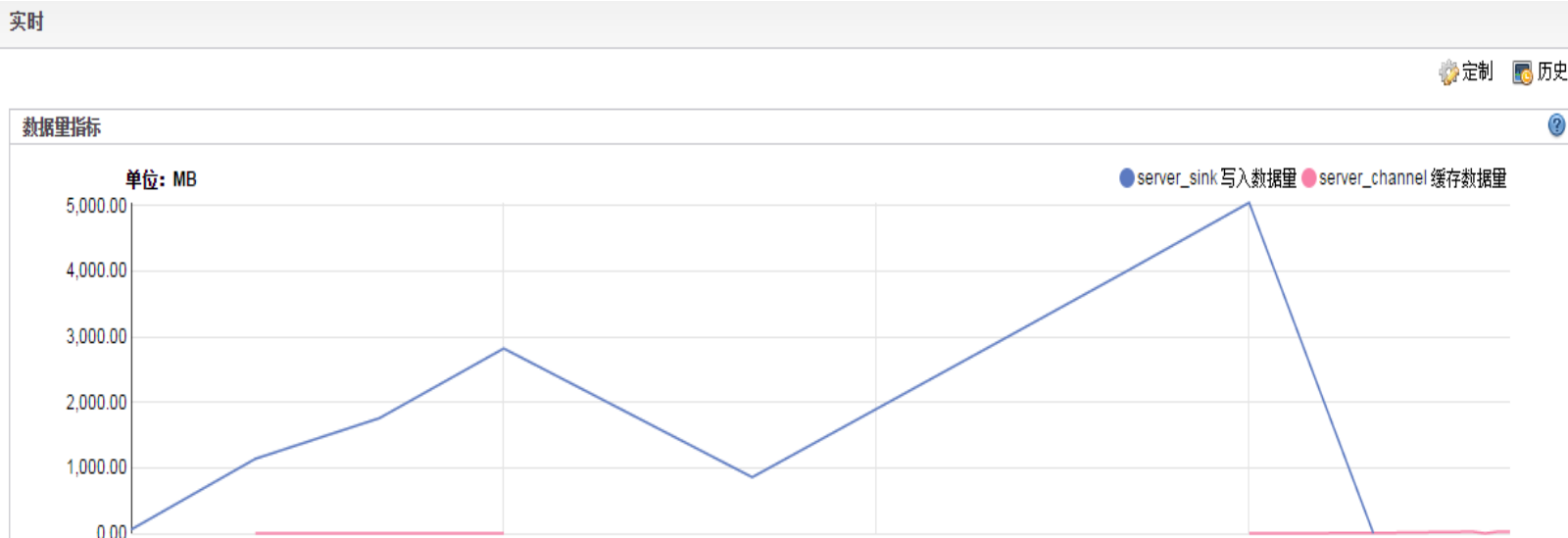
服务端**sink**修改的参数名和参数值为（其他参数值可以使用工具的默认值）：

SinkName	server_sink
type	kafka
kafka.topic	flume_test
channel	server_channel

点击“生成配置文件”按钮，将工具所在路径下的**properties.properties**配置文件上传到**192.168.1.100**这个节点的**Flume**实例上

配置运行结果

在节点192.168.1.20的/tmp/flume_test下放置一个采集文件，然后在Flume服务端192.168.1.100上观察监控指标数据变化，**server_sink**数据量在不断变化，表明写入数据成功，如图



思考题

- ① **Flume**采集数据特点
- ② **Flume**应用场景



本章总结

- 本章主要介绍了**Flume**应用场景，并对**Flume**的一些基本概念、配置工具使用做了详细说明。通过本章的学习，能够清楚知道**Flume**的作用、适用场景以及如何正确配置使用**Flume**。



习题

- 判断题
 1. **Flume**有客户端和服务端。 (T or F)
 2. 使用工具在配置**source/channel/sink**时必须先配置名字。 (T or F)
 3. **source/channel/sink** 的**type**不同时，相应的其他配置参数也不同。 (T or F)
- 单选题
 1. **Flume**当前不支持的**source**有 () ?
 - A.HDFS source
 - B.avro source
 - C.HTTP source
 - D.Kafka source

Thank you

www.huawei.com