

Flume组件安装与实践

Flume版本: 1.11.0

一、安装

(1) 解压缩

```
1 sudo tar -zxvf apache-flume-1.11.0-bin.tar.gz -C /usr/local
```

```
hadoop@hadoop-virtual-machine:/usr/local/sqoop/lib$ sudo tar -zxvf /mnt/hgfs/forShare/apache-flume-1.9.0-bin.tar.gz -C /usr/local
```

(2) 重命名和赋权限

```
1 sudo mv /usr/local/apache-flume-1.11.0-bin/ /usr/local/flume
2
3 sudo chown -R hadoop:hadoop /usr/local/flume
```

```
hadoop@hadoop-virtual-machine:/usr/local$ sudo mv apache-flume-1.9.0-bin/ flume
hadoop@hadoop-virtual-machine:/usr/local$ sudo chown hadoop:hadoop flume
```

(3) 把bin目录添加到path中:

```
1 $ gedit ~/.bashrc
```

写入下列内容:

```
1 export FLUME_HOME=/usr/local/flume
2 export PATH=$PATH:$FLUME_HOME/bin
```

```
1 $ source ~/.bashrc
```

(4) 修改flume-env.sh配置文件

```
1 cd /usr/local/flume/conf
2 cp ./flume-env.sh.template ./flume-env.sh
3 gedit ./flume-env.sh
```

写入以下内容：

```
1 export JAVA_HOME=/usr/local/jdk1.8.0_181
```

二、简单例子：从telnet接收数据，并在控制台输出

(1) 编写配置文件：

文件名：simple-example.conf

文件位置：/usr/local/flume/conf

文件内容：

```
1 a1.sources=r1
2 a1.sinks=k1
3 a1.channels=c1
4
5 a1.sources.r1.type=netcat
6 a1.sources.r1.bind=localhost
7 a1.sources.r1.port=44444
8
9 a1.sinks.k1.type=logger
10
11 a1.channels.c1.type=memory
12 a1.channels.c1.capacity=1000
13 #说明默认该通道中最大的可以存储的event数量是1000
```

```
14 a1.channels.c1.transactionCapacity=100
15 #说明每次最大可以从source中拿到或者送到sink中的event数量是100
16
17 a1.sources.r1.channels=c1
18 #说明数据源r1对应的通道是c1
19 a1.sinks.k1.channel=c1
20 #说明sinks k1对应的通道是c1
```

(2) 运行程序:

```
1 flume-ng agent --name a1 --conf-file /usr/local/flume/conf/simple-example.conf -
  Dflume.root.logger=INFO,console
```

(3) 启动telnet发送消息

打开另外一个终端，执行telnet命令，连接44444端口。

```
1 $ telnet localhost 44444
```

然后输入任意信息，如下：



```
hadoop@hadoop-virtual-machine:/usr/local/flume/conf$ telnet localhost 44444
Trying 127.0.0.1...
Connected to localhost.
Escape character is '^]'.
111
OK
222
OK
333
OK
444
OK
```

(4) 在启动flume任务的窗口查看消息，可以查看到对应接收到的信息：

```
hadoop@hadoop-virtual-machine: /usr/local/flume/conf
ce: Source starting
2023-09-11 20:51:29,910 INFO source.NetcatSource: Created serverSocket:sun.nio.ch.ServerSocketChannelImpl[/127.0.0.1:44444]
2023-09-11 20:55:43,978 INFO sink.LoggerSink: Event: { headers:{} body: 31 31 31 0D
                                     111. }
2023-09-11 20:55:43,978 INFO sink.LoggerSink: Event: { headers:{} body: 32 32 32 0D
                                     222. }
2023-09-11 20:55:43,978 INFO sink.LoggerSink: Event: { headers:{} body: 33 33 33 0D
                                     333. }
2023-09-11 20:55:43,979 INFO sink.LoggerSink: Event: { headers:{} body: 34 34 34 0D
                                     444. }
```

三、实例：从特定目录采集数据到HDFS

(1) 编写配置文件

#配置组件名称

```
1 a1.sources = r1
2 a1.sinks = k1
3 a1.channels = c1
4
5 #配置Source名称
6 a1.sources.r1.type = spooldir
7 a1.sources.r1.spoolDir = /usr/local/data/weblogs
8 a1.sources.r1.fileSuffix = .COMPLETED
9 a1.sources.r1.fileHeader = true
10 a1.sources.r1.ignorePattern = ([^ ]*.tmp)
11
12 #配置Sink组件
13 a1.sinks.k1.type = hdfs
14 a1.sinks.k1.hdfs.path = hdfs://192.168.101.130:9000/weblogs
15 a1.sinks.k1.hdfs.writeFormat = Text
16 a1.sinks.k1.hdfs.fileType = DataStream
17 a1.sinks.k1.hdfs.rollInterval = 600
18 a1.sinks.k1.hdfs.rollSize = 134217700
19 a1.sinks.k1.hdfs.rollCount = 0
20 a1.sinks.k1.hdfs.filePrefix = %Y-%m-%d-%H-%M-%S
21 a1.sinks.k1.hdfs.useLocalTimeStamp = true
```

```
22
23 #配置Channel组件
24 a1.channels.c1.type = memory
25 a1.channels.c1.capacity = 1000
26 a1.channels.c1.transactionCapacity = 100
27 # Source组件和Sink组件与Channel组件绑定
28 a1.sources.r1.channels = c1
29 a1.sinks.k1.channel = c1
```

(2) 创建文件夹

本地文件夹：/usr/local/data/weblogs

Hdfs文件夹：/weblogs

```
1 mkdir /usr/local/data/weblogs
2 hdfs dfs -mkdir /weblogs
```

```
hadoop@hadoop-virtual-machine:/usr/local/data$ mkdir weblogs
hadoop@hadoop-virtual-machine:/usr/local/data$ hdfs dfs -mkdir /weblogs
```

(3) 启动服务

```
1 flume-ng agent --name a1 --conf-file /usr/local/flume/conf/hdfspool.conf
```

(4) 复制文件到本地文件夹weblogs，并查看weblogs文件夹

```
1 cd /usr/local/data
2 cp test.txt weblogs
3 cp word1.txt weblogs
4 cp word2.txt weblogs
5 ls weblogs
```

```
hadoop@hadoop-virtual-machine:/usr/local/data$ cp test.txt weblogs
hadoop@hadoop-virtual-machine:/usr/local/data$ cp word1.txt weblogs
hadoop@hadoop-virtual-machine:/usr/local/data$ cp word2.txt weblogs
hadoop@hadoop-virtual-machine:/usr/local/data$ ls weblogs
test.txt.COMPLETED word1.txt.COMPLETED word2.txt.COMPLETED
hadoop@hadoop-virtual-machine:/usr/local/data$
```

被采集过的文件的后缀名都被改写为.COMPLETED了。

(5) flume服务窗口中有消息提醒

```
hadoop@hadoop-virtual-machine:/usr/local/data
2023-09-11 21:19:00,947 INFO avro.ReliableSpoolingFileEventRe
ader: Preparing to move file /usr/local/data/weblogs/word1.tx
t to /usr/local/data/weblogs/word1.txt.COMPLETED
2023-09-11 21:19:00,948 INFO hdfs.HDFSDataStream: Serializer
= TEXT, UseRawLocalFileSystem = false
2023-09-11 21:19:00,969 INFO hdfs.BucketWriter: Creating hdfs
://192.168.101.130:9000/weblogs/2023-09-11-21-19-00.169443834
0949.tmp
2023-09-11 21:19:01,015 INFO hdfs.HDFSDataStream: Serializer
= TEXT, UseRawLocalFileSystem = false
2023-09-11 21:19:01,037 INFO hdfs.BucketWriter: Creating hdfs
://192.168.101.130:9000/weblogs/2023-09-11-21-19-01.169443834
1016.tmp
2023-09-11 21:19:08,970 INFO avro.ReliableSpoolingFileEventRe
ader: Last read took us just up to a file boundary. Rolling t
o the next file, if there is one.
2023-09-11 21:19:08,970 INFO avro.ReliableSpoolingFileEventRe
ader: Preparing to move file /usr/local/data/weblogs/word2.tx
t to /usr/local/data/weblogs/word2.txt.COMPLETED
2023-09-11 21:19:08,970 INFO hdfs.HDFSDataStream: Serializer
= TEXT, UseRawLocalFileSystem = false
2023-09-11 21:19:08,994 INFO hdfs.BucketWriter: Creating hdfs
://192.168.101.130:9000/weblogs/2023-09-11-21-19-08.169443834
```

(6) 查看采集到hdfs系统中的数据

```
1 hdfs dfs -ls /weblogs
```

可以查看到hdfs上采集到的文件。

```
hadoop@hadoop-virtual-machine:/usr/local/data$ hdfs dfs -ls /weblogs
Found 6 items
-rw-r--r-- 1 hadoop supergroup 4 2023-09-11 21:18 /weblogs/2023-09-11
-21-18-52.1694438332936.tmp
-rw-r--r-- 1 hadoop supergroup 7 2023-09-11 21:18 /weblogs/2023-09-11
-21-18-55.1694438335077.tmp
-rw-r--r-- 1 hadoop supergroup 8 2023-09-11 21:19 /weblogs/2023-09-11
-21-19-00.1694438340949.tmp
-rw-r--r-- 1 hadoop supergroup 8 2023-09-11 21:19 /weblogs/2023-09-11
-21-19-01.1694438341016.tmp
-rw-r--r-- 1 hadoop supergroup 8 2023-09-11 21:19 /weblogs/2023-09-11
-21-19-08.1694438348971.tmp
-rw-r--r-- 1 hadoop supergroup 7 2023-09-11 21:19 /weblogs/2023-09-11
-21-19-09.1694438349030.tmp
```