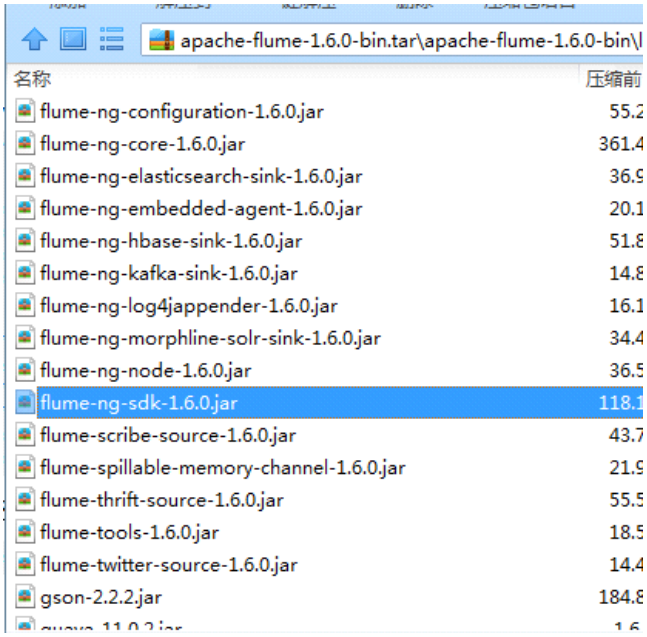
Flume_离线处理_日志收集

2019年3月10日 16:01

1. 发送日志到Flume

在日志服务器中，通过Log4jAppender将日志发往flume客户端

a. 在日志服务器应用中导入Log4jAppernder相关开发包



b. 配置log4j配置文件，实现发送日志给flume

```
log4j.rootLogger = info,stdout,flume


log4j.appender.stdout = org.apache.log4j.ConsoleAppender
log4j.appender.stdout.Target = System.out
log4j.appender.stdout.layout = org.apache.log4j.PatternLayout
log4j.appender.stdout.layout.ConversionPattern = %m%n


log4j.appender.flume = org.apache.flume.clients.log4jappender.Log4jAppender
log4j.appender.flume.Hostname = hadoop01
log4j.appender.flume.Port = 44444
log4j.appender.stdout.layout = org.apache.log4j.PatternLayout
log4j.appender.stdout.layout.ConversionPattern = %m%n
```

c. 在日志服务器的LogServlet中，通过log4j来发送日志

```
logger.info(line);
```

2. 开发客户端Agent

Hadoop03

```
#声明Agent
a1.sources = r1
a1.sinks = k1 k2
a1.channels = c1

#声明source
a1.sources.r1.type = avro
```

```
a1.sources.r1.bind = 0.0.0.0
a1.sources.r1.port = 44444

a1.sources.r1.interceptors = i1
a1.sources.r1.interceptors.i1.type = regex_extractor
a1.sources.r1.interceptors.i1.regex = ^(?:[^\\|]*\\|){14}\\d+_\\d+_(\\d+)\\|.*$
a1.sources.r1.interceptors.i1.serializers = s1
a1.sources.r1.interceptors.i1.serializers.s1.name = timestamp

#声明sink
a1.sinks.k1.type = avro
a1.sinks.k1.hostname = hadoop01
a1.sinks.k1.port = 44444

a1.sinks.k2.type = avro
a1.sinks.k2.hostname = hadoop02
a1.sinks.k2.port = 44444

a1.sinkgroups = g1
a1.sinkgroups.g1.sinks = k1 k2
a1.sinkgroups.g1.processor.type = load_balance
a1.sinkgroups.g1.processor.backoff = true
a1.sinkgroups.g1.processor.selector = random

#声明channel
a1.channels.c1.type = memory
a1.channels.c1.capacity = 1000
a1.channels.c1.transactionCapacity = 100

#绑定关系
a1.sources.r1.channels = c1
a1.sinks.k1.channel = c1
a1.sinks.k2.channel = c1
```

3. 开发中心服务器Agent

hadoop01 hadoop02

```
#配置Agent
a1.sources = r1
a1.sinks = k1
a1.channels = c1


#声明Source
a1.sources.r1.type = avro
a1.sources.r1.bind = 0.0.0.0
a1.sources.r1.port = 44444


#声明sink
a1.sinks.k1.type = hdfs
a1.sinks.k1.hdfs.path = hdfs://hadoop01:9000/flux/reportTime=%Y-%m-%d
a1.sinks.k1.hdfs.rollInterval = 30
a1.sinks.k1.hdfs.rollSize = 0
a1.sinks.k1.hdfs.rollCount = 0
a1.sinks.k1.hdfs.fileType = DataStream
a1.sinks.k1.hdfs.timeZone = GMT+8


#声明channel
```

```
#声明channel
a1.channels.c1.type = memory
a1.channels.c1.capacity = 1000
a1.channels.c1.transactionCapacity = 100


#绑定关系
a1.sources.r1.channels = c1
a1.sinks.k1.channel = c1
```

4. 遇到的问题

a. 找不到hadoop jar包

flume中的hdfs sink需要hadoop相关jar包的支持，

要么手动将hadoop相关jar包放置到flume的lib目录下

要么在本机中解压hadoop并将hadoop路径配置为HADOOP_HOME环境变量，使flume可以自动找到这些jar。

b. 产生大量小文件

hdfs sink的滚动条件设置不合理。

修改即可

```
a1.sinks.k1.hdfs.rollInterval = 30
a1.sinks.k1.hdfs.rollSize = 0
a1.sinks.k1.hdfs.rollCount = 0
```

c. 文件内容为乱码(序列化文件无法直接查看)

hdfs sink默认产生SequenceFile文件，无法直接查看

修改即可：

```
a1.sinks.k1.hdfs.fileType = DataStream
```

d. 希望能够按日期分目录存储

为了支持hive的分区处理，hdfs sink在将日志写入到hdfs的过程中，希望按照日期分目录存储。

```
a1.sinks.k1.hdfs.path = hdfs://hadoop01:9000/flux/reportTime=%Y-%m-%d
```

并且通过拦截器在日志头中增加timestamp头

```
a1.sources.r1.interceptors = i1
a1.sources.r1.interceptors.i1.type = regex_extractor
a1.sources.r1.interceptors.i1.serializers = s1
a1.sources.r1.interceptors.i1.serializers.s1.name = timestamp

```

e. 生成的目录时间不正确

配置hdfs采用的时区

```
a1.sinks.k1.hdfs.timeZone = GMT+8
```