Kafka实战－Flume到Kafka

1. 概述

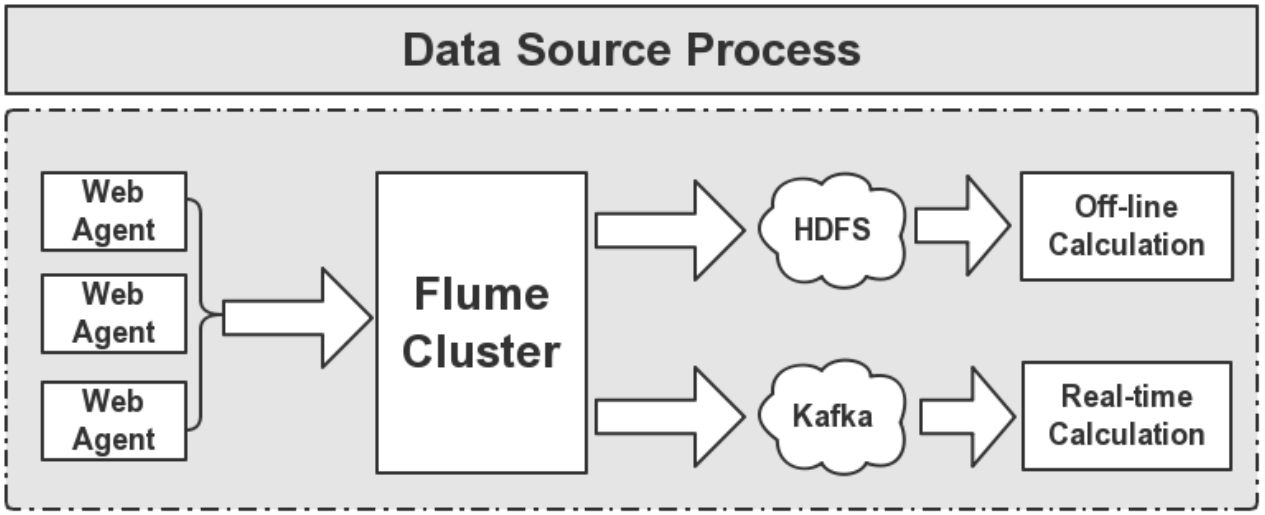
前面给大家介绍了整个Kafka项目的开发流程，今天给大家分享Kafka如何获取数据源，即Kafka生产数据。下面是今天要分享的目录：

- 数据来源
- Flume到Kafka
- 数据源加载
- 预览

下面开始今天的分享内容。

2. 数据来源

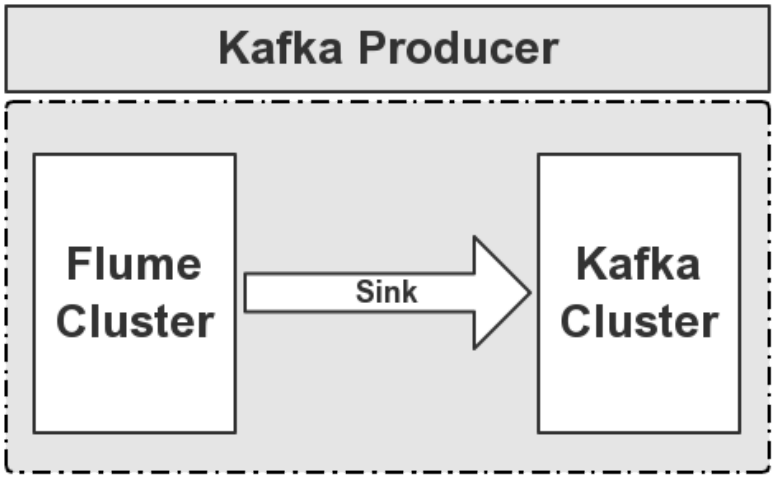
Kafka生产的数据，是由Flume的Sink提供的，这里我们需要用到Flume集群，通过Flume集群将Agent的日志收集分发到Kafka（供实时计算处理）和HDFS（离线计算处理）。关于Flume集群的Agent部署，这里就不多做赘述了，不清楚的同学可以参考《高可用Hadoop平台－Flume NG实战图解篇》一文中的介绍，下面给大家介绍数据来源的流程图，如下图所示：



这里，我们使用Flume作为日志收集系统，将收集到的数据输送到Kafka中间件，以供Storm去实时消费计算，整个流程从各个Web节点 上，通过Flume的Agent代理收集日志，然后汇总到Flume集群，在由Flume的Sink将日志输送到Kafka集群，完成数据的生产流程。

3.Flume到Kafka

从图，我们已经清楚了数据生产的流程，下面我们来看看如何实现Flume到Kafka的输送过程，下面我用一个简要的图来说明，如下图所示：



这个表达了从Flume到Kafka的输送工程，下面我们来看看如何实现这部分。

首先，在我们完成这部分流程时，需要我们将Flume集群和Kafka集群都部署完成，在完成部署相关集群后，我们来配置Flume的Sink数据流向，配置信息如下所示：

- 首先是配置spooldir方式，内容如下所示：

```
producer.sources.s.type = spooldir
producer.sources.s.spoolDir = /home/hadoop/dir/logdfs
```

- 当然，Flume的数据发送方类型也是多种类型的，有：Console、Text、HDFS、RPC等，这里我们系统所使用的是Kafka中间件来接收，配置内容如下所示：

```
1 producer.sinks.r.type = org.apache.flume.plugins.KafkaSink
2 producer.sinks.r.metadata.broker.list=dn1:9092,dn2:9092,dn3:9092
3 producer.sinks.r.partition.key=0
4 producer.sinks.r.partitioner.class=org.apache.flume.plugins.SinglePartition
```

```
producer.sinks.r.serializer.class=kafka.serializer.StringEncoder
producer.sinks.r.request.required.acks=0
producer.sinks.r.max.message.size=1000000
producer.sinks.r.producer.type=sync
producer.sinks.r.custom.encoding=UTF-8
producer.sinks.r.custom.topic.name=test
```

这样，我们就在Flume的Sink端配置好了数据流向接受方。

4. 数据加载

在完成配置后，接下来我们开始加载数据，首先我们在Flume的spooldir端生产日志，以供Flume去收集这些日志。然后，我们通过Kafka的KafkaOffsetMonitor监控工具，去监控数据生产的情况，下面我们开始加载。

- 启动ZK集群，内容如下所示：

```
zkServer.sh start
```

注意：分别在ZK的节点上启动。

- 启动Kafka集群

```
kafka-server-start.sh config/server.properties &
```

在其他的Kafka节点输入同样的命令，完成启动。

- 启动Kafka监控工具

```
java -cp KafkaOffsetMonitor-assembly-0.2.0.jar \
 com.quantifind.kafka.offsetapp.OffsetGetterWeb \
 --zk dn1:2181,dn2:2181,dn3:2181 \
 --port 8089 \
 --refresh 10.seconds \
 --retain 1.days
```

- 启动Flume集群

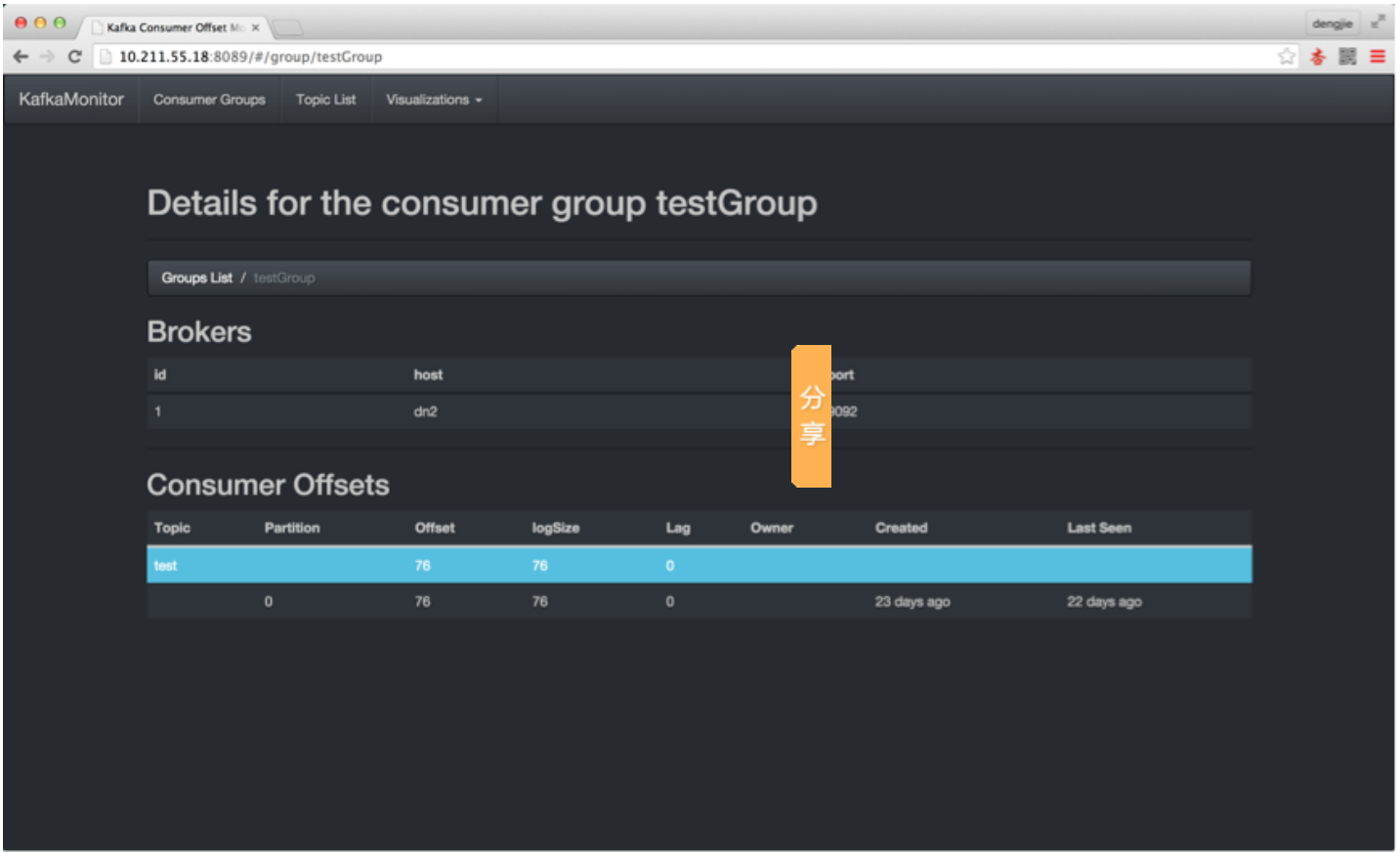
```
flume-ng agent -n producer -c conf -f flume-kafka-sink.properties -Dflume.root.logger=ERROR,console
```

然后，我在/home/hadoop/dir/logdfs目录下上传log日志，这里我只抽取了一少部分日志进行上传，如下图所示，表示日志上传成功。

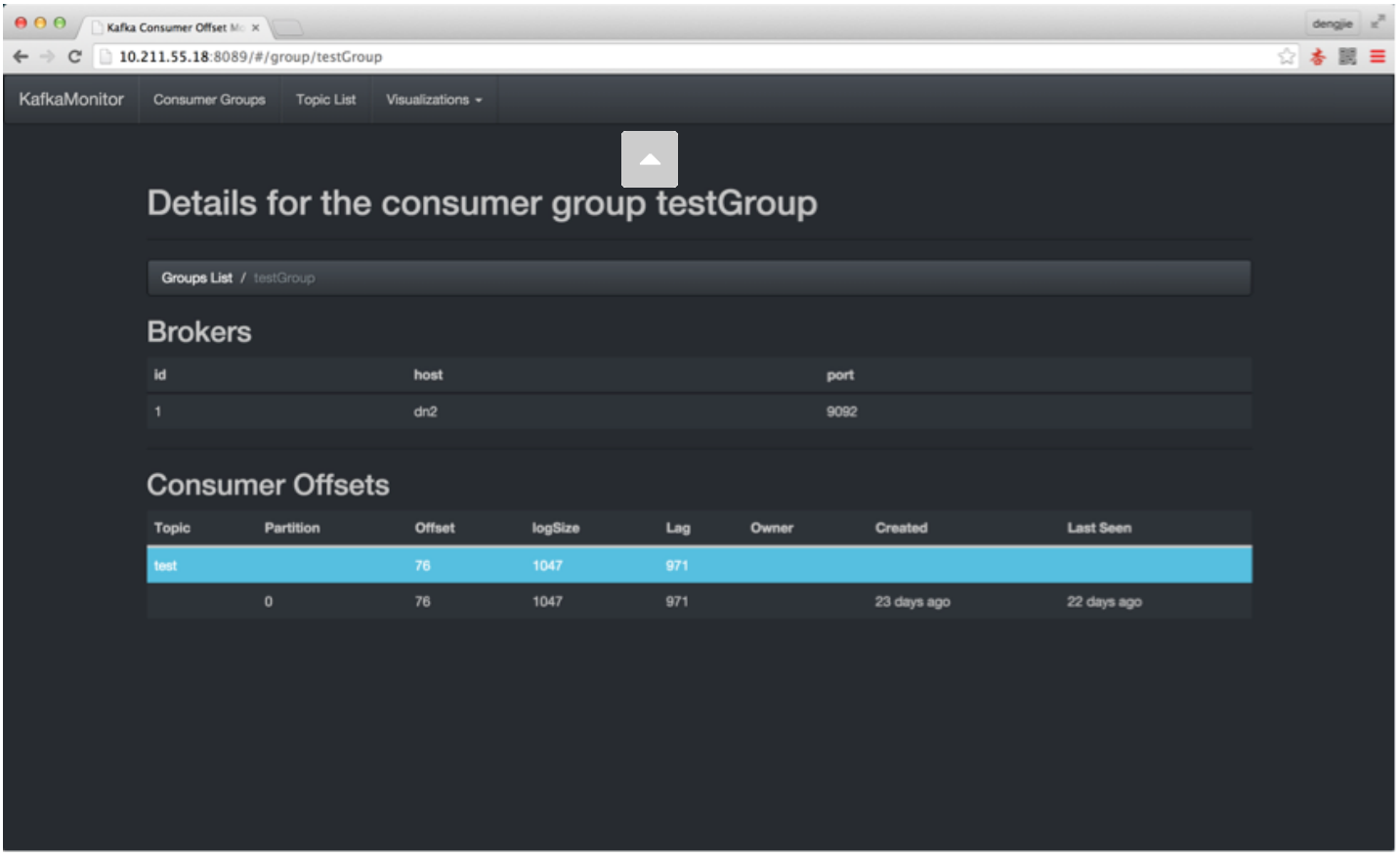
5. 预览

下面，我们通过Kafka的监控工具，来预览我们上传的日志记录，有没有在Kafka中产生消息数据，如下所示：

- 启动Kafka集群，为生产消息截图预览



- 通过Flume上传日志，在Kafka中产生消息数据



6. 总结

本篇文章给大家讲述了Kafka的消息产生流程，后续会在Kafka实战系列中为大家讲述Kafka的消息消费流程等一整套流程，这里只是为后续的Kafka实战编码打下一个基础，让大家先对Kafka的消息生产有个整体的认识。

来自：http://www.cnblogs.com/smartloli/p/4615908.html