## 需求分析

结构示意图:

采集需求：某服务器的某特定目录下，会不断产生新的文件，每当有新文件出现，就需要把文件采集到 HDFS 中去

根据需求，首先定义以下 3 大要素

- 数据源组件，即 source ——监控文件目录 ： spooldir

  spooldir 特性：

  1、监视一个目录，只要目录中出现新文件，就会采集文件中的内容

  2、采集完成的文件，会被 agent 自动添加一个后缀：COMPLETED

  3、所监视的目录中不允许重复出现相同文件名的文件
- 下沉组件，即 sink——HDFS 文件系统 ： hdfs sink
- 通道组件，即 channel——可用 file channel 也可以用内存 channel

## flume 配置文件开发

配置文件编写:

```
cd   /export/servers/apache-flume-1.6.0-cdh5.14.0-bin/conf
mkdir -p /export/servers/dirfile
vim spooldir.conf
```

```
# Name the components on this agent
a1.sources = r1
a1.sinks = k1
a1.channels = c1
# Describe/configure the source
##name is only one
a1.sources.r1.type = spooldir
a1.sources.r1.spoolDir = /export/servers/dirfile
a1.sources.r1.fileHeader = true
# Describe the sink
a1.sinks.k1.type = hdfs
a1.sinks.k1.channel = c1
a1.sinks.k1.hdfs.path                                                     =
hdfs://node01:8020/spooldir/files/%y-%m-%d/%H%M/
```

```
a1.sinks.k1.hdfs.filePrefix = events-
a1.sinks.k1.hdfs.round = true
a1.sinks.k1.hdfs.roundValue = 10
a1.sinks.k1.hdfs.roundUnit = minute
a1.sinks.k1.hdfs.rollInterval = 3
a1.sinks.k1.hdfs.rollSize = 20
a1.sinks.k1.hdfs.rollCount = 5
a1.sinks.k1.hdfs.batchSize = 1
a1.sinks.k1.hdfs.useLocalTimeStamp = true
#gen filestyle,default Sequencefile,use DataStream text
a1.sinks.k1.hdfs.fileType = DataStream
# 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
```

Channel 参数解释：
capacity：默认该通道中最大的可以存储的 event 数量
trasactionCapacity：每次最大可以从 source 中拿到或者送到 sink 中的 event 数量
keep-alive：event 添加到通道中或者移出的允许时间

## 启动 flume

```
bin/flume-ng agent -c ./conf -f ./conf/spooldir.conf -n a1 -Dflume.root.logger=INFO,console
```

## 上传文件到指定目录

将不同的文件上传到下面目录里面去，注意文件不能重名

```
cd /export/servers/dirfile
```