Tutorial 4 - Trabalhando um pouco com Flume

Para trabalharmos com o Flume temos que criar o arquivo de configuração do agente. O arquivo de configuração inclui as propriedades do *source*, *channel* e *sink* do agente e como eles são conectados para formar o fluxo de dados.

Cada componente (*source*, *channel* e *sink*) no fluxo possui um nome, um tipo e um conjunto de propriedades que são específicos para o tipo. Todos esses atributos de um componente precisam ser definidos no arquivo de propriedades do agente Flume.

Vamos ver um exemplo.

Objetivo: Nosso objetivo nesse tutorial é criar um fluxo de dados entre uma pasta em nosso sistema de arquivos local e o HDFS.

Primeiro vamos criar duas pastas dentro da pasta local testebigdata: dados e conf. Usaremos a pasta conf para salvar nosso agente Flume e a pasta dados para construir o canal com o HDFS.

No caso do HDFS vamos conectar com a pasta testeflume. Por isso, você deve criar essa pasta no Hadoop. Você pode criar por meio do comando:

hdfs dfs -mkdir testeflume

Abra o gedit e crie um arquivo chamado agente1.conf e salve na pasta testebigdata/conf (lembre-se: crie a pasta conf no testebigdata).

1. Vamos primeiro criar o agente e seu respectivo source, sink e channel.

```
<Agent>.sources = <Source>
<Agent>.sinks = <Sink>
<Agent>.channels = <Channel>
```

Em nosso exemplo, escreve no agente1.conf:

```
a1.sources = r1
a1.sinks = k1
a1.channels = c1
```

Cada componente tem sua configuração específica, mas em todas é fundamental o campo *type*.

2. source

```
<Agente>.sources.<Source>.type =
```

Em nosso exemplo vamos usar o tipo spooldir e indica qual a pasta em nosso diretório local vamos "ouvir"

```
a1.sources.r1.type = spooldir
a1.sources.r1.spoolDir = testebigdata/dados
```

3. sinks

Crie um novo diretório no HDFS (hdfs dfs -mkdir testeflume). Agora vamos definir que nosso sink armazenará os dados no HDFS e passaremos a pasta onde os dados devem ficar:

```
a1.sinks.k1.type = hdfs
a1.sinks.k1.hdfs.path = testeflume
```

4. channel

Aqui devemos informar qual o tipo do channel, o único campo requerido. As demais configurações variam de acordo com o tipo escolhido. Por exemplo, o tipo memory indica que os dados serão armazenados em memória. Podemos então definir a capacity que correspondem ao número máximo de eventos armazenados no canal (por padrão 100).

```
a1.channels.c1.type = memory
a1.channels.c1.capacity = 1000
```

O channel também pode ser do tipo file em que os dados ficam armazenados no arquivo local do agente. Também temos a opção do *Spillable Memory Channel* que funciona como um canal de memória até estar cheio. Nesse ponto, ele age como um canal de arquivo.

5. Ligando os componentes

Por fim, devemos ligar o source e sink ao channel.

```
a1.sources.r1.channels = c1
a1.sinks.k1.channel = c1
```

Como executar o agente?

Um agente é iniciado o comando *flume-ng* que está localizado no diretório bin da distribuição Flume. Você precisa especificar o agente nome, o diretório de configuração e o arquivo de configuração na linha de comando:

```
$ flume-ng agent --conf <diretório> --conf-file <caminho do arquivo>
--name <nome do agente>
```

Vamos executar o exemplo que criamos.

```
$ flume-ng agent --conf testebigdata/conf --conf-file
testebigdata/conf/agente1.conf --name a1
```

Relembrando nosso agente:

```
a1.sources = r1
a1.sinks = k1
a1.channels = c1

a1.sources.r1.type = spooldir
a1.sources.r1.spoolDir = testebigdata/dados

a1.sinks.k1.type = hdfs
a1.sinks.k1.hdfs.path = testeflume

a1.channels.c1.type = memory
a1.channels.c1.capacity = 1000

a1.sources.r1.channels = c1
a1.sinks.k1.channel = c1
```

Assim um canal será criado entre a pasta local testebigdata/dados e a pasta testeflume que está no HDFS. E tudo que for inserido na pasta dados será transferido para o testeflume enquanto o agente estiver em execução no terminal.