Hadoop: Hands-On: Flume and ZooKeeper

FLUME CONFIGURATION

Configuration:

```
# example.conf: A single-node Flume configuration

# Name the components on this agent
  a1.sources = r1
a1.sinks = k1
  a1.channels = c1

# Describe/configure the source
  a1.sources.r1.type = netcat
  a1.sources.r1.bind = localhost
  a1.sources.r1.port = 44444

# Describe the sink
  a1.sinks.k1.type = logger

# 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
```

Given this configuration file, start Flume as follows:

```
$ bin/flume-ng agent --conf conf --conf-file example.conf --name a1 -
  Dflume.root.logger=INFO,console
```

Here is an example of enabling both configuration logging and raw data logging while setting the Log4j log level to DEBUG for the console output:

```
$ bin/flume-ng agent --conf conf --conf-file example.conf --name a1 -
  Dflume.root.logger=DEBUG,console -Dorg.apache.flume.log.printconfig=true -
  Dorg.apache.flume.log.rawdata=true
```

Starting an agent:

An agent is started using a shell script called 'flume-ng' which is located in the bin directory of the Flume distribution. You need to specify the agent name, the config directory, and the config file on the command line:

```
$ bin/flume-ng agent -n $agent_name -c conf -f conf/flume-conf.properties.template
```

ZOOKEEPER-BASED CONFIGURATION

Flume supports agent configurations via ZooKeeper. This is an experimental feature. The configuration file needs to be uploaded in ZooKeeper, under a configurable prefix. The configuration file is stored in the ZooKeeper Node data. Following is how the ZooKeeper Node tree would look like for agents, a1 and a2:

```
- /flume
|- /a1 [Agent config file]
|- /a2 [Agent config file]
```

Once the configuration file is uploaded, start the agent with following options:

```
$ bin/flume-ng agent –conf conf -z zkhost:2181,zkhost1:2181 -p /flume –name a1 -
Dflume.root.logger=INFO,console
Argument Name         DefaultDescription
z        –       Zookeeper connection string. Comma separated list of hostname:port
p        /flume  Base Path in Zookeeper to store Agent configurations
```

For example, an agent named 'agent_foo' is reading data from an external Avro client and sending it to HDFS via a memory channel. The config file weblog.config would look like the below:

```
# list the sources, sinks and channels for the agent
  agent_foo.sources = avro-appserver-src-1
  agent_foo.sinks = hdfs-sink-1
agent_foo.channels = mem-channel-1

# set channel for source
agent_foo.sources.avro-appserver-src-1.channels = mem-channel-1

# set channel for sink
agent_foo.sinks.hdfs-sink-1.channel = mem-channel-1
```

Configuring individual components:

```
# properties for sources
<Agent>.sources.<Source>.<someProperty> = <someValue>

# properties for channels
<Agent>.channel.<Channel>.<someProperty> = <someValue>

# properties for sinks
<Agent>.sources.<Sink>.<someProperty> = <someValue>
```