Lab: Importing Log Data into HDFS using Flume

About this Lab

Objective: Import data from a log file into HDFS using Flume.

File locations: /root/devph/labs/Lab3.3

Successful outcome: The data in webtraffic.log on sandbox will be streamed into HDFS
directory
/user/root/flumedata/.

Before you begin: Your HDP 2.3 cluster should be up and running within your VM.

Related lesson: *Flume*

Lab Steps

1) Verify Flume is installed on your cluster.

    a. If not already done, open a Terminal in your VM and type "ssh sandbox".
    b. Type the following command from prompt:

```
# flume-ng
```

    c. The command should return the following usage instructions:

```
[root@node1 ~]# flume-ng
Error: Unknown or unspecified command ''

Usage: /usr/hdp/2.2.0.0-2041/flume/bin/flume-ng.distro <command> [options]...

commands:
  help                  display this help text
  agent                 run a Flume agent
  avro-client           run an avro Flume client
  password              create a password file for use in flume config
  version               show Flume version info

global options:
  --conf,-c <conf>      use configs in <conf> directory
  --classpath,-C <cp>   append to the classpath
  --dryrun,-d           do not actually start Flume, just print the command
  --plugins-path <dirs> colon-separated list of plugins.d directories. See the
                        plugins.d section in the user guide for more details.
                        Default: $FLUME_HOME/plugins.d
  -Dproperty=value      sets a Java system property value
  -Xproperty=value      sets a Java -X option

agent options:
  --conf-file,-f <file> specify a config file (required)
```

2) View the contents of the webtraffic.log file

       a. Change directories to /root/devph/labs/Labs3.3.

```
# cd /root/devph/labs/Lab3.3
```

       b. View the contents of the webtraffic.log:

```
# less webtraffic.log
```

  3) Press q to quit viewing the log.

       a. A partial agent configuration has been written for you, view the file using the less command.

```
# less logagent.conf
```

       b. From the VM desktop, start gedit.
       c. Click Open and navigate to /root/devph/labs/Lab3.3
       d. Modify this file replacing strings starting with "REPLACE-WITH-" with their appropriate value. HINT: Search for "HDFS Sink" at https://flume.apache.org/FlumeUserGuide.html.

```
agent.sources = weblog
agent.channels = memoryChannel
agent.sinks = mycluster

## Sources
##################################################
agent.sources.weblog.type = exec
agent.sources.weblog.command = tail -F REPLACE-WITH-PATH2-
webtraffic.log-FILE
agent.sources.weblog.batchSize = 1
agent.sources.weblog.channels = REPLACE-WITH-CHANNEL-NAME

## Channels
##################################################
agent.channels.memoryChannel.type = memory
agent.channels.memoryChannel.capacity = 100
agent.channels.memoryChannel.transactionCapacity = 100

## Sinks
###################################################
agent.sinks.mycluster.type = REPLACE-WITH-CLUSTER-TYPE
agent.sinks.mycluster.hdfs.path=/user/root/flumedata
agent.sinks.mycluster.channel = REPLACE-WITH-CHANNEL-NAME
```

e. Click Save to save the logagent configuration.
f. From the terminal window, start the logagent from /root/devph/labs/Lab3.3.

```
# flume-ng agent –n agent –f logagent.conf &
```

g. Hit Enter to get a command prompt back.
h. The sink for logagent is the HDFS folder /user/root/flumedata. Verify that this folder does now exists.

```
# hdfs dfs –ls flumedata/
```

i. View the contents of the streamed data in HDFS in FlumeData.<########>.

```
# hdfs dfs –cat flumedata/FlumeData.<######>
```

The flumedata/FlumeData.<sequence> file should display the webtraffic as shown below:

Notice the format difference.

Result

Successful import of data from Flume into HDFS.