Working with Flume

ABOUT FLUME

Apache Flume is a tool/service/data ingestion mechanism for collecting aggregating and transporting large amounts of streaming data such as log files, events (etc...) from various sources to a centralized data store.

Flume is a highly reliable, distributed, and configurable tool. It is principally designed to copy streaming data (log data) from various web servers to HDFS.

CASE SCENARIO

Now we will use Flume with source type **SpoolDir** . We can ingest data by placing files to be ingested into a "spooling" directory on disk with this "spooling" directory. Moreover, though, this Apache Flume source we will watch the specified directory for new files and will parse events out of new files as they appear. However, the event parsing logic is pluggable. Also, it renames to indicate completion after a given file will fully read into the channel.
In addition, a Spooling Directory Flume source is reliable and will not miss data. Although, even if Flume is restarted or killed, unlike the Exec source.

In this case we will create a directory called **testdata** on Linux and sink directory on hdfs called **flumetest**. And will create configuration file in flume where source will be testdata and sink will be flumetest. The source type will be spooldir which means as new file arrives in testdata (source), flume will read and push into flumetest (sink). It will mark the file as .completed in source directory once it is pushed.

STEP 1

Create directory called testdata on Linux

mkdir testdata

STEP 2

Create sink directory on hdfs called flumetest

#**hdfs dfs - mkdir flumetest**

STEP 3

Now we need to create configuration file of Flume defining agent, source channel and sink

$ **sudo vi /etc/hadoop/flumespool.conf**

Add below lines and save the file

```
agent1.sinks = hdfs_sink
agent1.sources = spool_source
agent1.channels = mem_channel

agent1.channels.mem_channel.type = memory
agent1.channels.mem_channel.capacity = 500


# Define a source for agent1
agent1.sources.spool_source.type = spooldir
agent1.sources.spool_source.spoolDir = /home/training/testdata
agent1.sources.spool_source.fileHeader = false
agent1.sources.spool_source.fileSuffix = .COMPLETED


# Defining sink for agent1
agent1.sinks.hdfs_sink.type = hdfs
agent1.sinks.hdfs_sink.hdfs.path = /user/training/flumetest
agent1.sinks.hdfs_sink.hdfs.batchSize = 1000
agent1.sinks.hdfs_sink.hdfs.rollSize = 268435456
agent1.sinks.hdfs_sink.hdfs.rollInterval = 0
agent1.sinks.hdfs_sink.hdfs.rollCount = 50000000
agent1.sinks.hdfs_sink.hdfs.writeFormat=Text
agent1.sinks.hdfs_sink.hdfs.fileType = DataStream


agent1.sources.spool_source.channels = mem_channel
agent1.sinks.hdfs_sink.channel = mem_channel
```

Press ESC key, and type :wq to save and exit.

Important description of properties mentioned in file

fileType - This is the required file format of our HDFS file. SequenceFile, DataStream and CompressedStream are the three types available with this stream. In our example, we are using the DataStream.

writeFormat - Could be either text or writable.

batchSize - It is the number of events written to a file before it is flushed into the HDFS. Its default value is 100.

rollsize - It is the file size to trigger a roll. It default value is 100.

rollCount - It is the number of events written into the file before it is rolled. Its default value is 10.

STEP 4

Start the agent on new terminal and you must see the message source started, leave it running.

```
# sudo flume-ng agent --conf-file /etc/hadoop/flumespool.conf \
  --name agent1
```

STEP 5

Now open a new terminal and create 2 files with data1 and data2 names by adding lines in that as shown below

$ **vi data1**

Press ESC and type i key for insert and type below line

I am learning hadoop

Press ESC and type :wq to save and exit

$ **vi data2**

Press ESC and type i key for insert and type below line

I am learning spool in flume

Press ESC and type :wq to save and exit

STEP 6

Now we will copy data1 and data2 file into testdata

cp data1 testdata (enter)

Now go to the terminal where you started the agent as in Step 4 and observe the message at last

Repeat the same for data2

cp data2 testdata (enter)

STEP 7

Now press CTRL + C (to stop the agent) in terminal where agent is running, which was started in Step 4.

STEP 8

Check the hdfs

hdfs dfs -ls /user/training/flumetest

You will find a file created by name something like this: FlumeData.1561472586448

Now cat this file and you will observe that data1 and data2 file data (2 lines) will be shown, it means it is spooling all the files data from source to single file on sink.

hdfs dfs -cat /user/training/flumetest/FlumeData.1561472586448

CASE 2 (ONLY FOR REFERENCE DO NOT EXECUTE)

Twitter example:

 # Naming the components on the current agent.

TwitterAgent.sources = Twitter

TwitterAgent.channels = MemChannel

TwitterAgent.sinks = HDFS

 # Describing/Configuring the source

TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource

TwitterAgent.sources.Twitter.consumerKey = Your OAuth consumer key

TwitterAgent.sources.Twitter.consumerSecret = Your OAuth consumer secret

TwitterAgent.sources.Twitter.accessToken = Your OAuth consumer key access token

TwitterAgent.sources.Twitter.accessTokenSecret = Your OAuth consumer key access token secret

TwitterAgent.sources.Twitter.keywords = java, bigdata, mapreduce

 # Describing/Configuring the sink

TwitterAgent.sinks.HDFS.type = hdfs

TwitterAgent.sinks.HDFS.hdfs.path = hdfs://elephant:8020/user/Hadoop/twitter_data/

TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream

TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text

TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000

TwitterAgent.sinks.HDFS.hdfs.rollSize = 0

TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000

 # Describing/Configuring the channel

TwitterAgent.channels.MemChannel.type = memory

TwitterAgent.channels.MemChannel.capacity = 10000

TwitterAgent.channels.MemChannel.transactionCapacity = 100

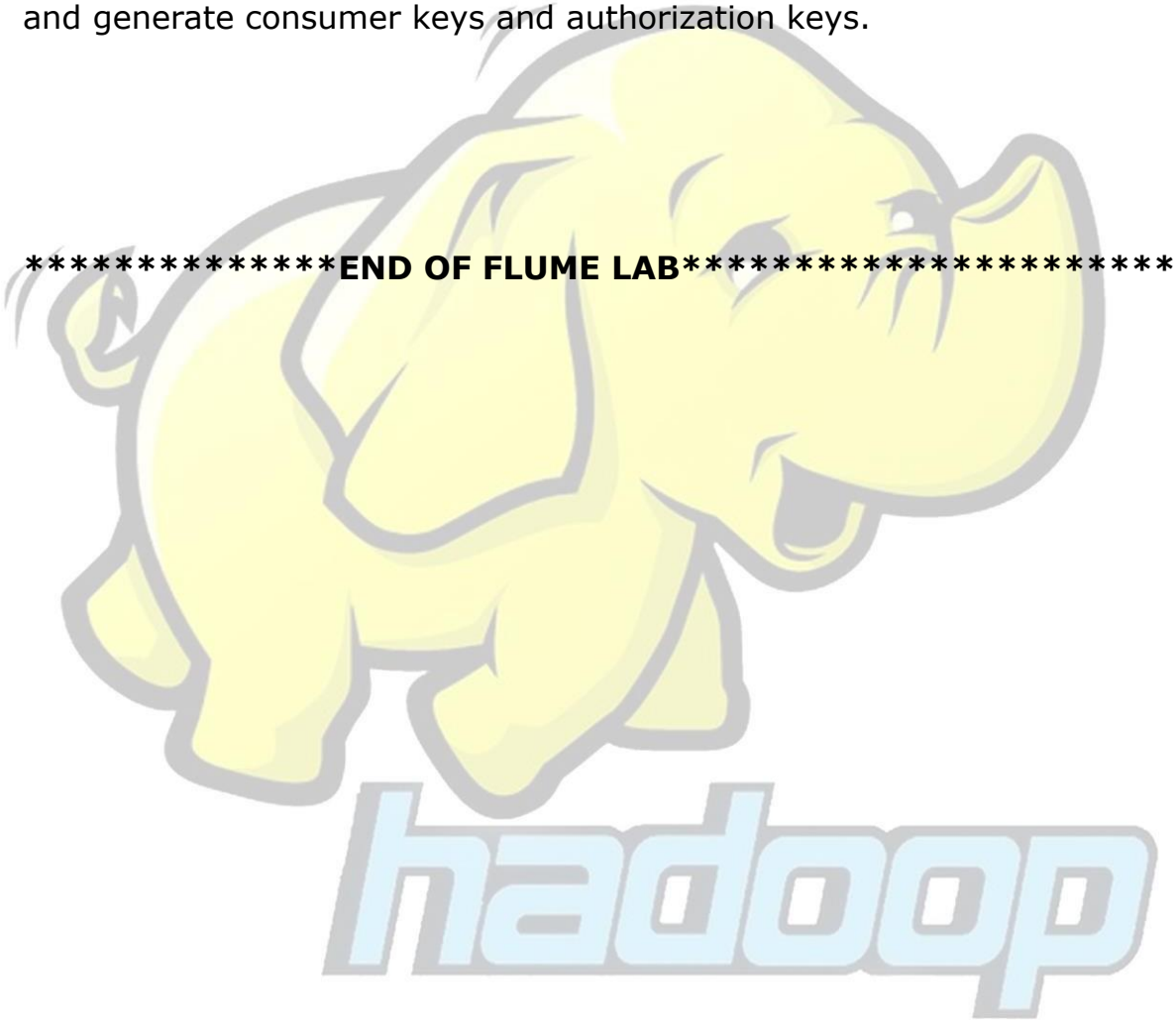
 # Binding the source and sink to the channel

TwitterAgent.sources.Twitter.channels = MemChannel

TwitterAgent.sinks.HDFS.channel = MemChannel

NOTE

You have to create twitter account , and twitter API in twitter developer site and generate consumer keys and authorization keys.



****************END OF FLUME LAB*************************