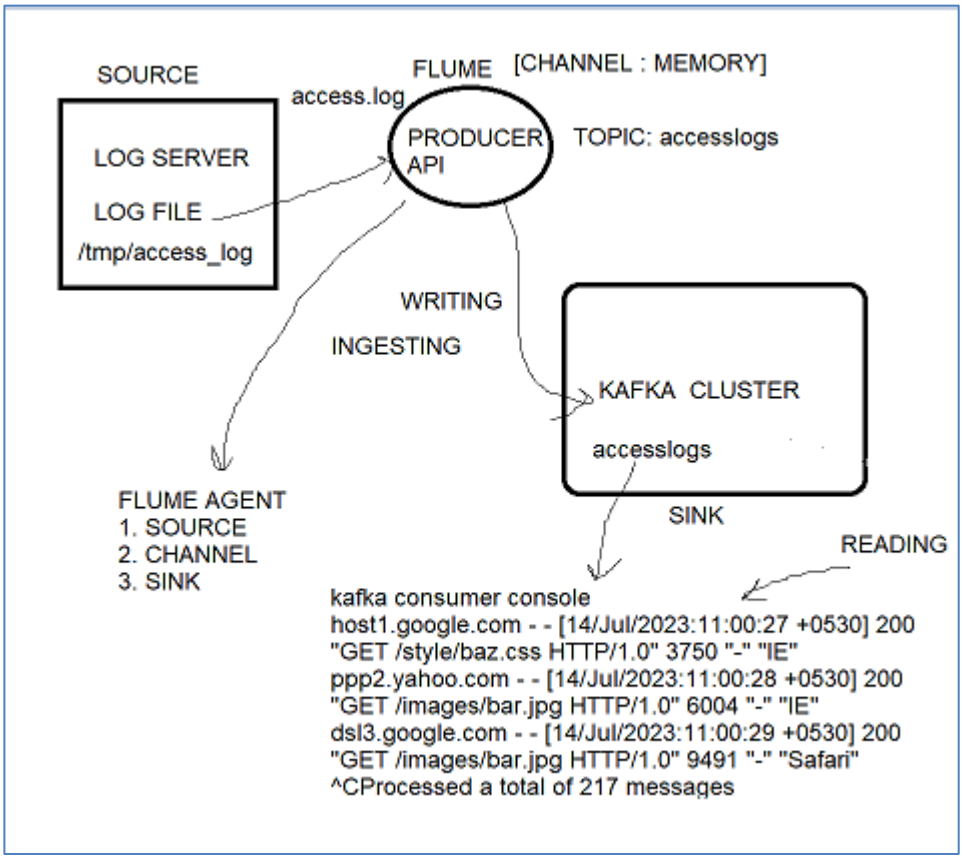
FLUME AND KAFKA [DATA PIPELINE]

In this scenario we are going to practically implement a use case to demonstrate how Flume and Kafka can integrate.

Flume will be used as Producer API to read logs [Files] from local filesystem, and copy into the Kafka topic.

We will read contents of the log files ingested by Flume as messages using Kafka consumer API.



STEP 1

Install flume and set path.

1. Download the flume tar file from the drive and copy it into Home folder [Present on Desktop] of Elephant machine.
2. Extract the tar file.

tar -xvf apache-flume-1.11.0-bin.tar.gz

mv apache-flume-1.11.0-bin /root/flume

cd [Press Enter]

vi .bashrc

Add the below at last of the file.

export FLUME_HOME=/root/flume
export PATH=$PATH:$FLUME_HOME/bin/

Save the file.

. .bashrc

flume-ng -help

Must show the options.

STEP 2

Create flume configuration file.

cd /root/flume

vi flume-conf.properties

Add the below lines

```
flagent.sources = src1
flagent.channels = ch1
flagent.sinks = sink1

flagent.sources.src1.type = exec
flagent.sources.src1.command = tail -F /tmp/access_log
flagent.sources.src1.channels = ch1

flagent.channels.ch1.type = memory
flagent.channels.ch1.capacity = 500

flagent.sinks.sink1.type = org.apache.flume.sink.kafka.KafkaSink
flagent.sinks.sink1.topic = accesslogs
flagent.sinks.sink1.brokerList = elephant.training.com:9092
flagent.sinks.sink1.requiredAcks = 1
flagent.sinks.sink1.batchSize = 20
flagent.sinks.sink1.channel = ch1
```

Save the file.

STEP 3

Create topic in kafka by name accesslogs.

kafka-topics --create --bootstrap-server elephant.training.com:9092 --replication-factor 1 --partitions 1 --topic accesslogs

STEP 4

Now open 3 terminals

On first terminal start the kafka consumer console

kafka-console-consumer --bootstrap-server elephant.training.com:9092 --topic accesslogs

Leave this terminal as it is since it is waiting for the messages.

On second terminal start the bash script to generate the log files.

Copy the accesslog-gen.bash on Elephant machine desktop

cd Desktop

chmod 777 accesslog-gen.bash

./accesslog-gen.bash /tmp/access_log

Leave the terminal and go to another terminal

On third terminal Start the flume agent

flume-ng agent --conf-file /root/flume/flume-conf.properties --name flagent

```
[root@saibigdata ~]# flume-ng agent --conf-file /root/flume/flume-conf.properties --name flagent
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hadoop libraries found via (/bin/hadoop) for HDFS access
Info: Including Hive libraries found via () for Hive access
+ exec /usr/java/jdk1.8.0_211-amd64/bin/java -Xmx20m -cp '/root/flume/lib/*:/etc/hadoop/conf:/usr/lib/hadoop/lib/*:/usr/lib/hadoop/.//*:/
adoop-hdfs/./:/usr/lib/hadoop-hdfs/lib/*:/usr/lib/hadoop-hdfs/.//*:/usr/lib/hadoop-mapreduce/.//*:/usr/lib/hadoop-yarn/lib/*:/usr/lib/had
.//*:/lib/*' -Djava.library.path=:/usr/lib/hadoop/lib/native org.apache.flume.node.Application --conf-file /root/flume/flume-conf.propert
e flagent
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/root/flume/lib/log4j-slf4j-impl-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
```

Now leave the flume terminal and go to the terminal where you have started kafka consumer console

```
[root@saibigdata Desktop]# kafka-console-consumer --bootstrap-server saibigdata.training.com:9092 --topic accesslogs
ppp2.yahoo.com - - [14/Jul/2023:12:03:00 +0530] 200 "GET /style/baz.css HTTP/1.0" 4320 "-" "Mozilla"
host1.google.com - - [14/Jul/2023:12:03:01 +0530] 200 "GET /style/baz.css HTTP/1.0" 5578 "-" "IE"
ppp2.msn.com - - [14/Jul/2023:12:03:02 +0530] 200 "GET /images/bar.jpg HTTP/1.0" 10091 "-" "IE"
dsl3.google.com - - [14/Jul/2023:12:03:03 +0530] 200 "GET /docs/foo.html HTTP/1.0" 4421 "-" "Safari"
dsl3.google.com - - [14/Jul/2023:12:03:04 +0530] 200 "GET /docs/foo.html HTTP/1.0" 272 "-" "Mozilla"
dsl3.google.com - - [14/Jul/2023:12:03:05 +0530] 200 "GET /docs/foo.html HTTP/1.0" 5846 "-" "Firefox"
host1.google.com - - [14/Jul/2023:12:03:06 +0530] 200 "GET /docs/foo.html HTTP/1.0" 2860 "-" "IE"
host1.google.com - - [14/Jul/2023:12:03:07 +0530] 200 "GET /images/bar.jpg HTTP/1.0" 5838 "-" "Mozilla"
host1.msn.com - - [14/Jul/2023:12:03:08 +0530] 200 "GET /docs/foo.html HTTP/1.0" 401 "-" "Safari"
host1.yahoo.com - - [14/Jul/2023:12:03:09 +0530] 200 "GET /style/baz.css HTTP/1.0" 4953 "-" "Mozilla"
ppp2.google.com - - [14/Jul/2023:12:03:10 +0530] 200 "GET /style/baz.css HTTP/1.0" 123 "-" "Firefox"
ppp2.yahoo.com - - [14/Jul/2023:12:03:11 +0530] 200 "GET /images/bar.jpg HTTP/1.0" 3054 "-" "Firefox"
ppp2.yahoo.com - - [14/Jul/2023:12:03:12 +0530] 200 "GET /images/bar.jpg HTTP/1.0" 7833 "-" "IE"
```

We can see the messages coming!

***** Happy Learning *****