Flume document by Mr.Sasidhar
Cloudera Certified Hadoop professional
Sasis937@gmail.com

Flume Material

Overview

Apache Flume is a distributed, reliable, and available system for efficiently collecting, aggregating and moving large amounts of log data from many different sources to a centralized data store.

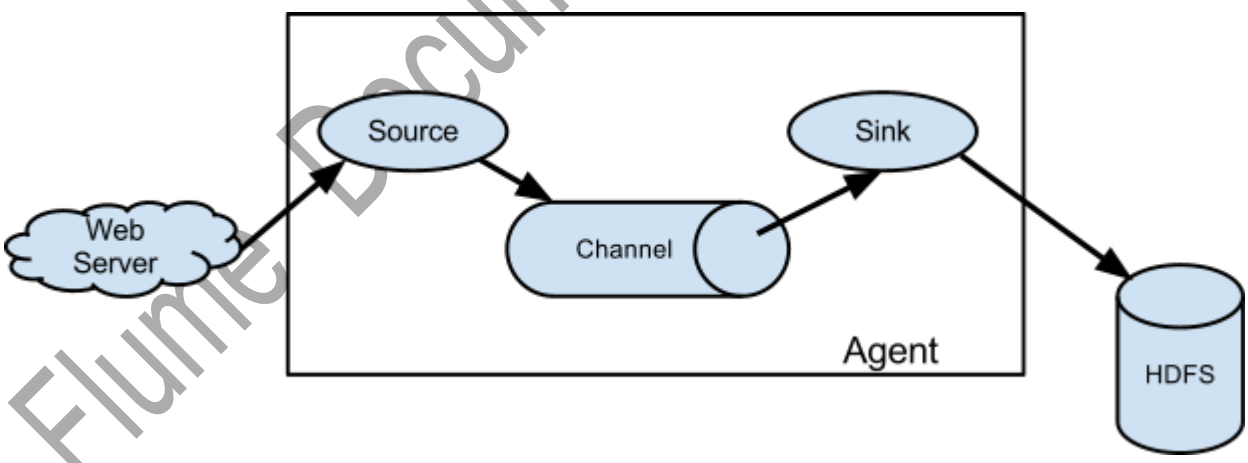
System Requirements

1. Java Runtime Environment - Java 1.6 or later (Java 1.7 Recommended)
2. Memory - Sufficient memory for configurations used by sources, channels or sinks
3. Disk Space - Sufficient disk space for configurations used by channels or sinks
4. Directory Permissions - Read/Write permissions for directories used by agent

Architecture

Data flow model

A Flume event is defined as a unit of data flow having a byte payload and an optional set of string attributes. A Flume agent is a (JVM) process that hosts the components through which events flow from an external source to the next destination (hop).



Practical process

1. The first step is to create an application in https://dev.twitter.com/apps/ and then generate the corresponding keys.

2) Assuming that Hadoop has already been installed and configured, the next step isdownload Flume (Already software shared) and extract it to any folder through tar -xzvf command. Then specify the Environmental variables.
chmod 777 apache-flume…..
tar –xzvf apache-flume…..

Environmental variables (gedit .bashrc)
export FLUME_HOME=/home/sbkt/apache-flume-1.6.0-bin
export PATH=$PATH:$FLUME_HOME/bin
source .bashrc // For reflecting Environmental variables

3) Specify your $JAVA_HOME as shown below in the conf/flume-env.sh file
export JAVA_HOME=/usr/java/jdk1.6.0_39

4) Download the flume-sources-1.0-SNAPSHOT.jar (Already shared software) and add it to the flume class path as shown below in the conf/flume-env.sh file

```
FLUME_CLASSPATH="/home/training/Installations/apache-flume-1.3.1-
                bin/flume-sources-1.0-SNAPSHOT.jar"
```

5)The conf/flume.conf should have all the agents (flume, memory and hdfs) defined as below

```
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS

TwitterAgent.sources.Twitter.type =
com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey = <consumerKey>
TwitterAgent.sources.Twitter.consumerSecret = <consumerSecret>
TwitterAgent.sources.Twitter.accessToken = <accessToken>
TwitterAgent.sources.Twitter.accessTokenSecret = <accessTokenSecret>

TwitterAgent.sources.Twitter.keywords = hadoop, big data, analytics

TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path =
hdfs://localhost:9000/user/flume/tweets/
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
```

```
TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100
```

The **consumerKey**, **consumerSecret**, **accessToken** and **accessTokenSecret** have to be replaced with those obtained from https://dev.twitter.com/apps.
And, **TwitterAgent.sinks.HDFS.hdfs.path** should point to the NameNode and the location in HDFS where the tweets will go to.

The **TwitterAgent.sources.Twitter.keywords** value can be modified to get the tweets for some other topic like football, movies etc

5) Start flume using the below command

flume-ng agent -n TwitterAgent -c conf -f /home/sbkt/apache-flume-1.6.0-bin/conf/flume.conf

After a couple of minutes the Tweets should appear in your specified directory (hdfs://localhost:9000/user/flume/tweets/) in HDFS.

If any errors, feel free to contact
Mr. Sasidhar
Sasis937@gmail.com