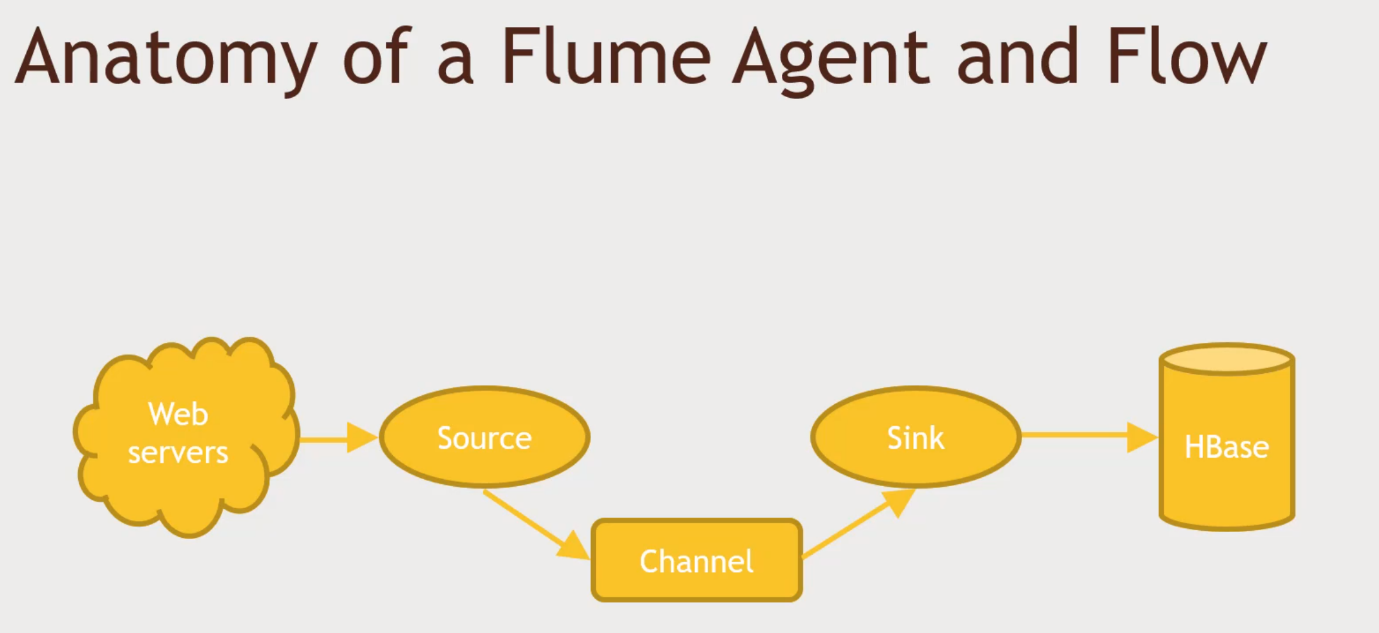
**Flume** More fun with data streaming

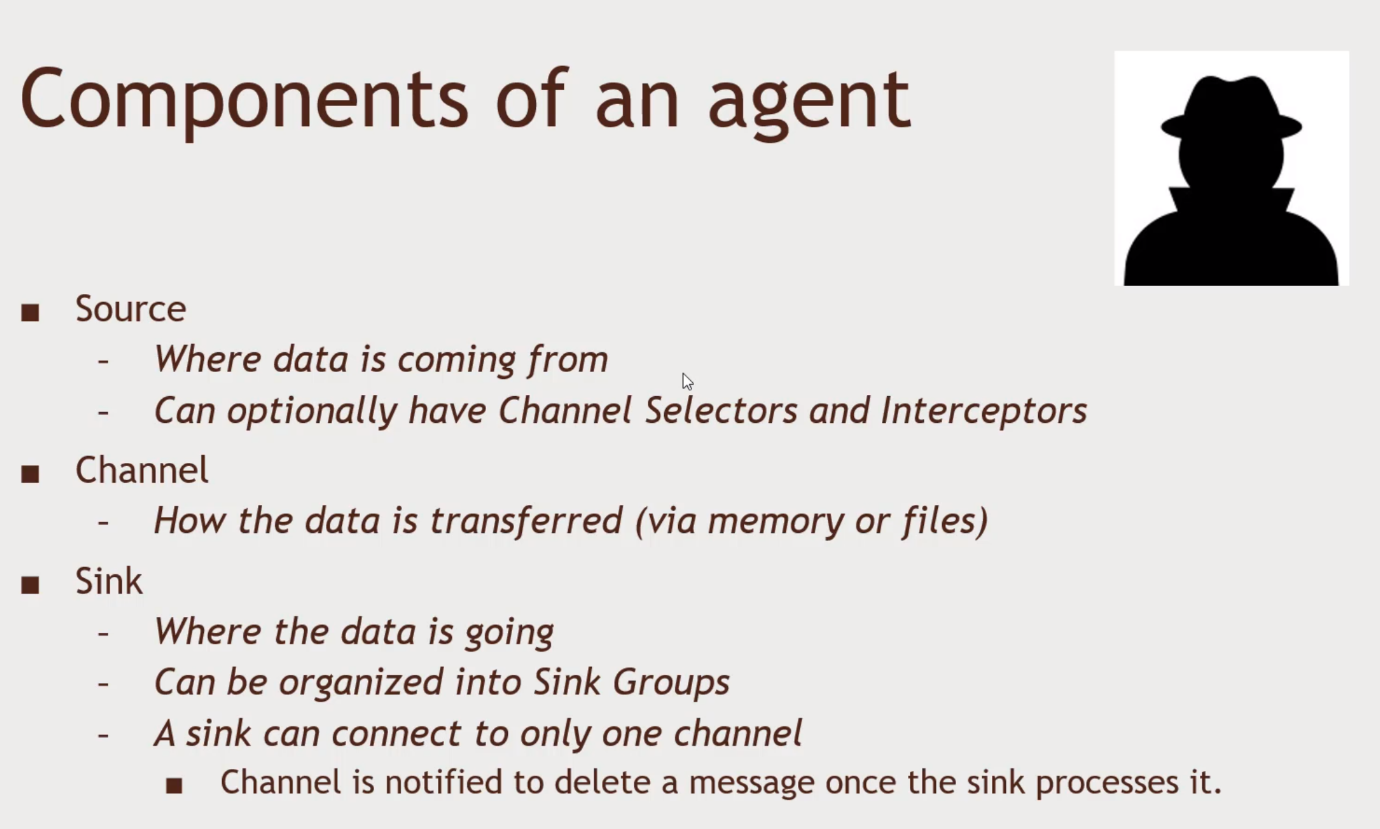
Another way to stream data into your cluster

Made from the start with Hadoop in mind

* Built in sinks for HDFS and H base

Originally made to handle log aggregation

Source->listens to webserver logs

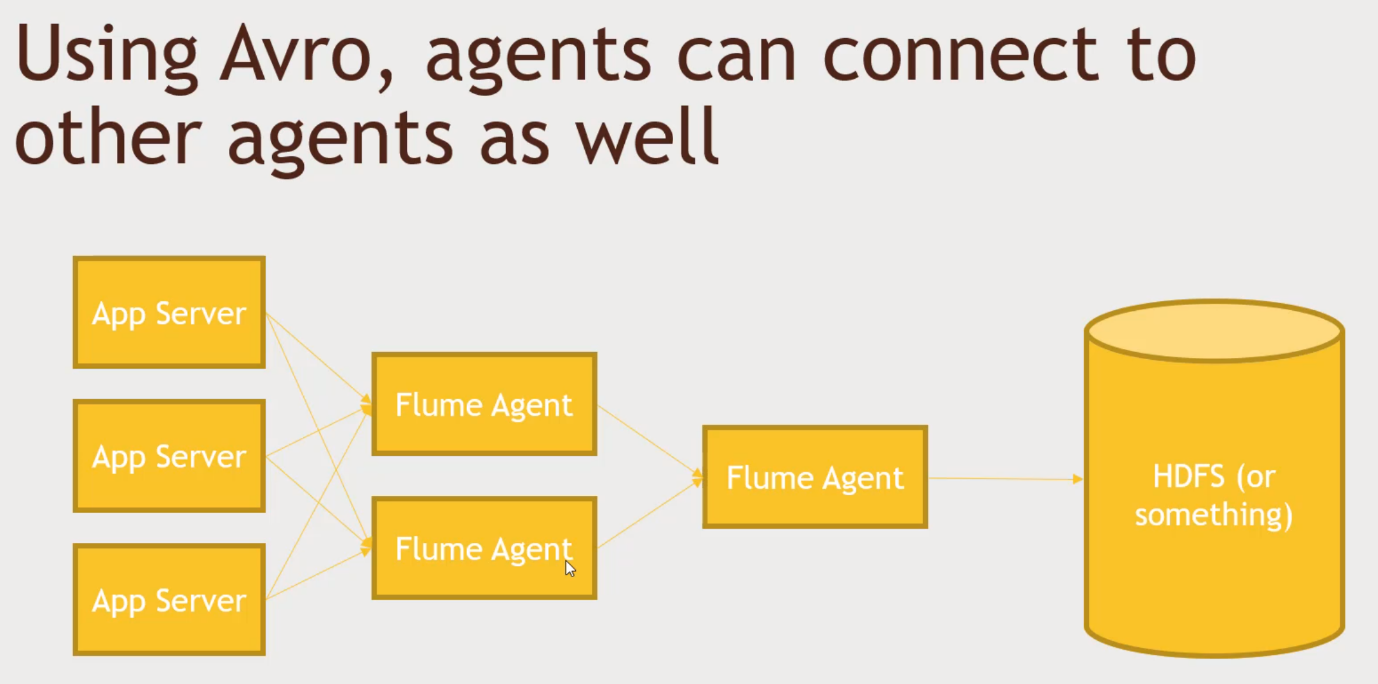
Memory/file Channel Transfer into Sink, write into HBase or HDFS

Bulit-in Source Types

* Spooling Directory
* Avro
* Kafka
* Exec – Command Line interface
* Thrift – Data connection interface
* Netcat -Listen to data being streamed in from TCP port web traffic
* HTTP
* Custom

Built-in Sink Types

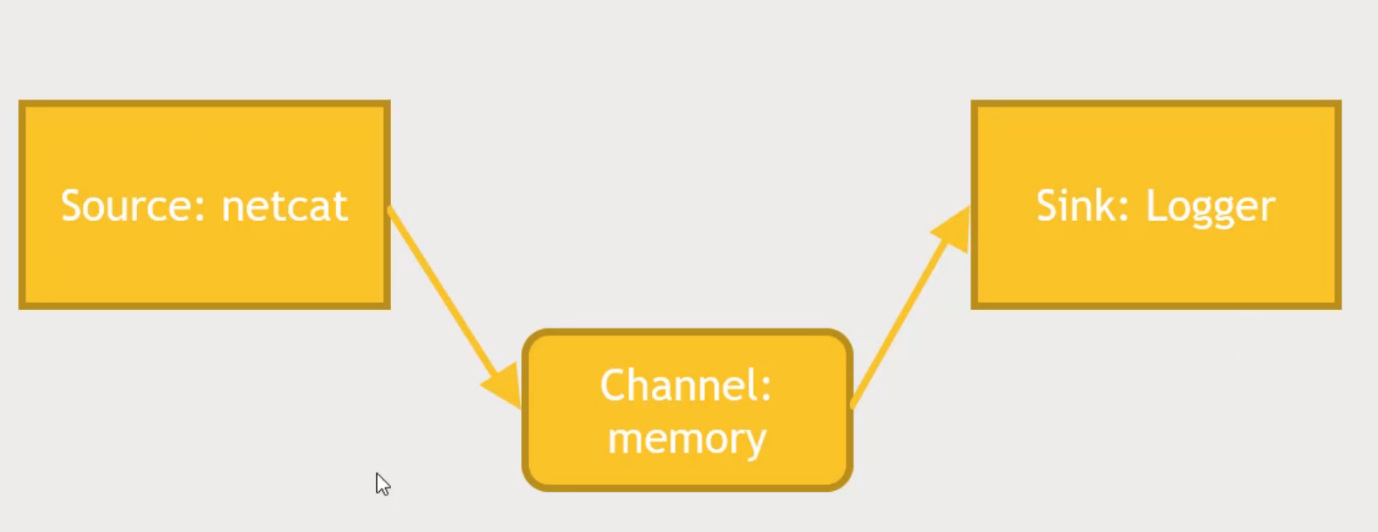
* HDFS
* Hive
* HBase
* Avro
* Thrift
* Elasticsearch
* Kafka
* Custom
* And more!

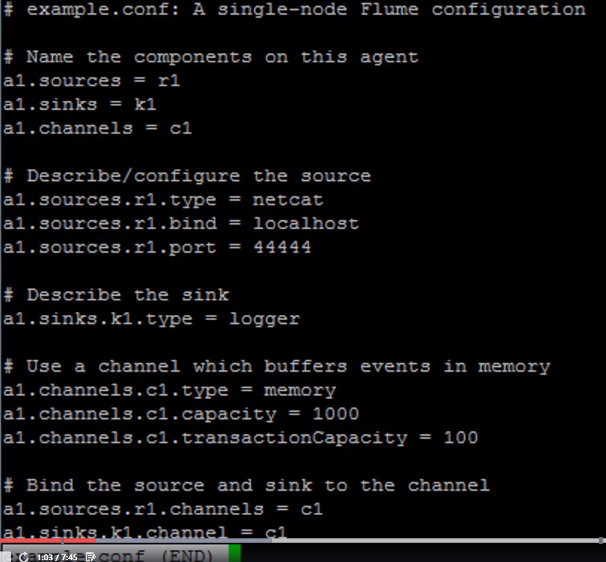
Using Avro

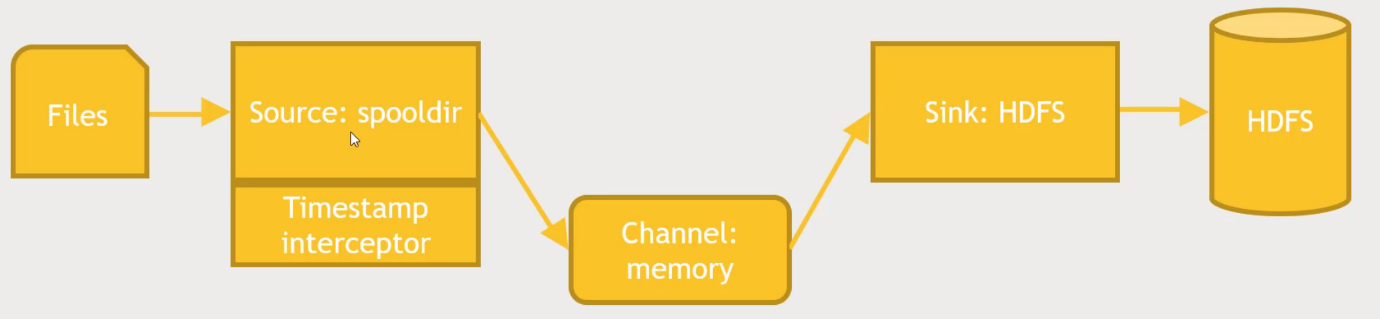
Avro data procotol Avro is gluing agent

Think of Flume as a buffer between your data and your cluster. Smoothing out the traffic between the data source and the cluster.

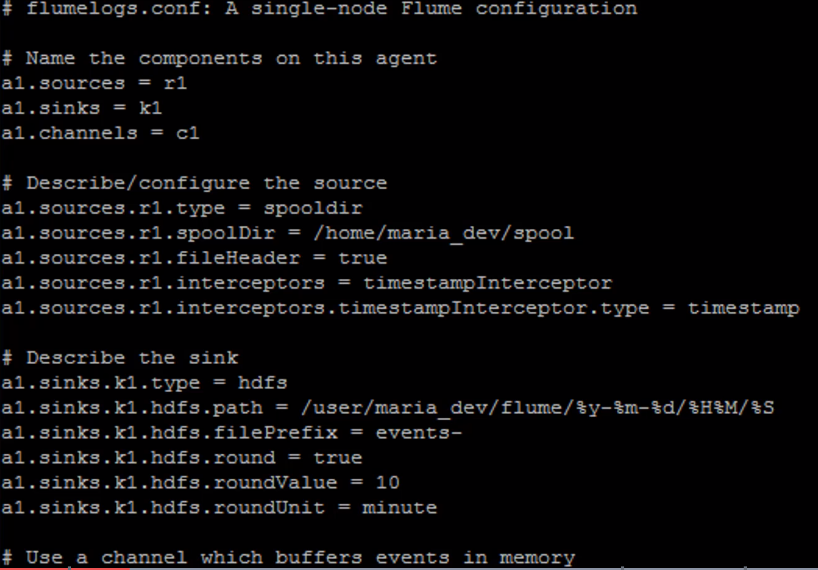
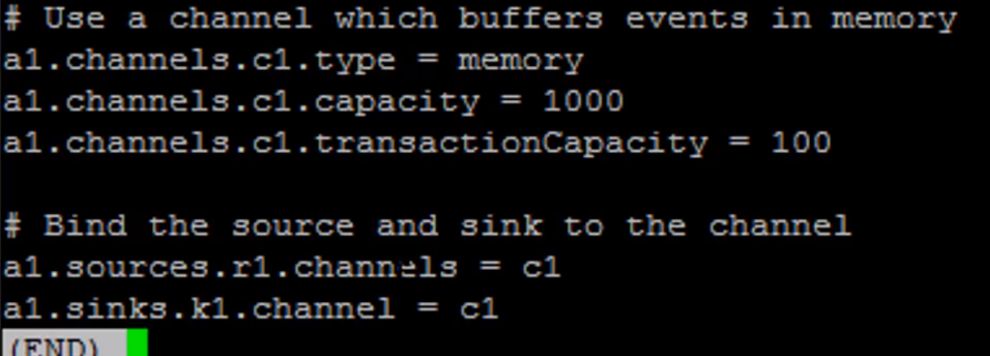
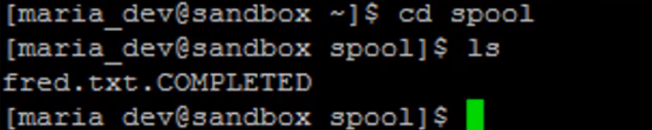
Flume is built in in our HortonWorks. Practice

Net cat listens to TCP port for data.

1. Set up the Hortonwork sandbox and login to Putty command prompt
2. Write a config file, sources, channel and sink you want in your setup
   1. wget media.sundog-soft.com/hadoop/example.conf
   2. less example.conf (Sink can only have 1 channel)
3. Open up a new Putty session
   1. Start up flume
      1. cd /usr/hdp/current/flume-server/
      2. bin/flume-ng agent --conf conf --conf-file ~example.conf --name a1 -Dflume.root.logger=INFO,console
4. Go back to the previous putty session
   1. Type in the following
      1. telnet localhost 44444
      2. Hello there, how are you today. It will be logged in the /var/log/flume/flume.log.file.
      3. Ctrl + right ] to logout

Log spools to HDFS. Have Flume monitor for new files that are dropped into it and put those in HDFS in a certain file structure.

2 Sessions.

1. Home directory of Maria\_dev
   1. wget media.sundog-soft.com/Hadoop/flumelogs.conf
   2. less flumelogs.conf
   3. It has time interceptor to get the timestamp to flow through the sink from the channel. Monitor a spooldir for new file, if a new file is found and go through every line and send data to the sink. Sink is in HDFS which will store the file. Round the minutes down to a 10 min interval.
   4. Connect all the dots together.
2. Create the directory that we are going to be using
   1. mkdir spool
   2. Bring up Ambari and Login as admin
      1. Files View -user/maria\_dev and new folder -> flume
3. Start the flume server in another putty prompt.
   1. bin/flume-ng agent --conf conf --conf-file ~flumelogs.conf --name a1 -Dflume.root.logger=INFO,console
4. Test if its working
   1. cp access\_log\_small.txt spool/fred.txt
   2. then you should see something happens.
   3. cd spool
5. Under Flume directory
   1. There will be a folder with Date, Time, and you will find all of the event’s file. And se one line. Its in SEQ file format.