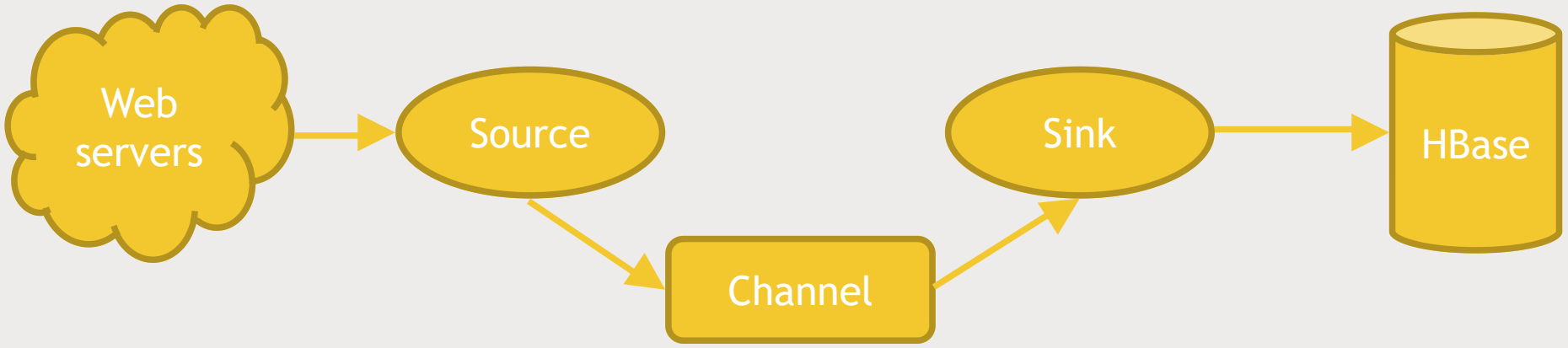
FLUME

More fun with data streaming

What is Flume?

- Another way to stream data into your cluster
- Made from the start with Hadoop in mind
  - *Built-in sinks for HDFS and Hbase*
- Originally made to handle log aggregation

Anatomy of a Flume Agent and Flow



Components of an agent

- Source
  - *Where data is coming from*
  - *Can optionally have Channel Selectors and Interceptors*
- Channel
  - *How the data is transferred (via memory or files)*
- Sink
  - *Where the data is going*
  - *Can be organized into Sink Groups*
  - *A sink can connect to only one channel*
    - Channel is notified to delete a message once the sink processes it.

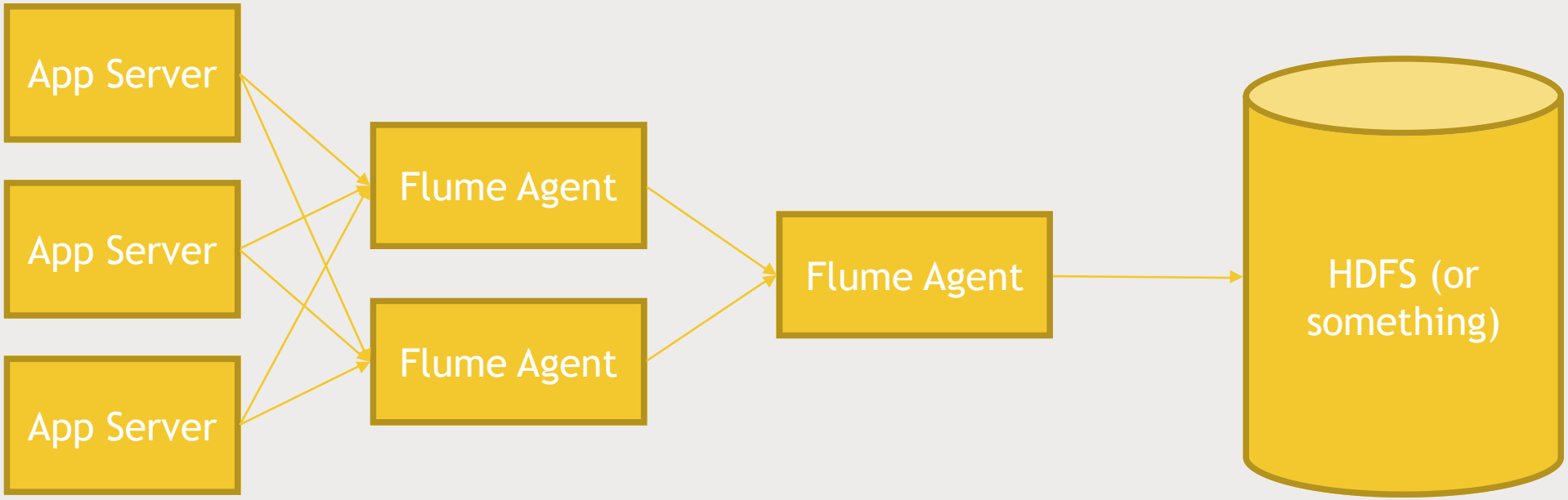
Built-in Source Types

- Spooling directory
- Avro
- Kafka
- Exec
- Thrift
- Netcat
- HTTP
- Custom
- And more!

Built-in Sink Types

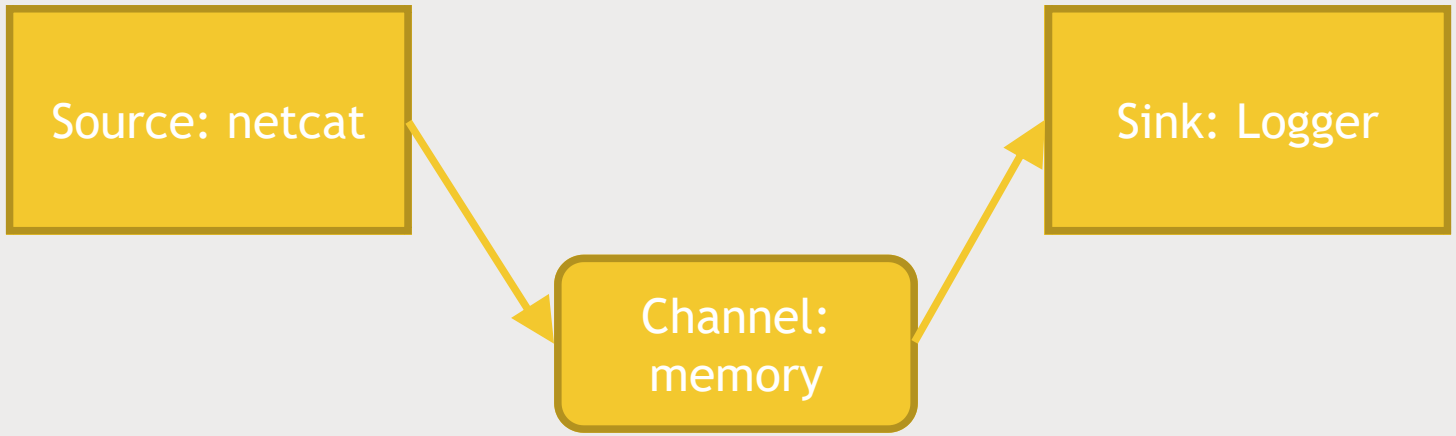
- HDFS
- Hive
- HBase
- Avro
- Thrift
- Elasticsearch
- Kafka
- Custom
- And more!

Using Avro, agents can connect to other agents as well



Think of Flume as a buffer between your data and your cluster.

Let’s play: Simple flow



Let’s play: log spool to HDFS