

Apache Flume

Data Collection System

Agenda

- Overview of Flume
- Flume Sources
- Channels & Sinks
- Flume Topology
- Production Architecture
- Monitoring & Performance



- Collection, Aggregation of streaming Event Data
 - ❖ Typically used for log data
- Significant advantages over ad-hoc solutions
 - ❖ Reliable, Scalable, Manageable, Customizable and High Performance
 - ❖ Declarative, Dynamic Configuration
 - ❖ Contextual Routing
 - ❖ Feature rich
 - ❖ Fully extensible



An Event is the fundamental unit of data transported by Flume from its point of origination to its final destination. Event is a byte array payload accompanied by optional headers.

- Payload is opaque to Flume
- Headers are specified as an unordered collection of string key-value pairs, with keys being unique across the collection
- Headers can be used for contextual routing

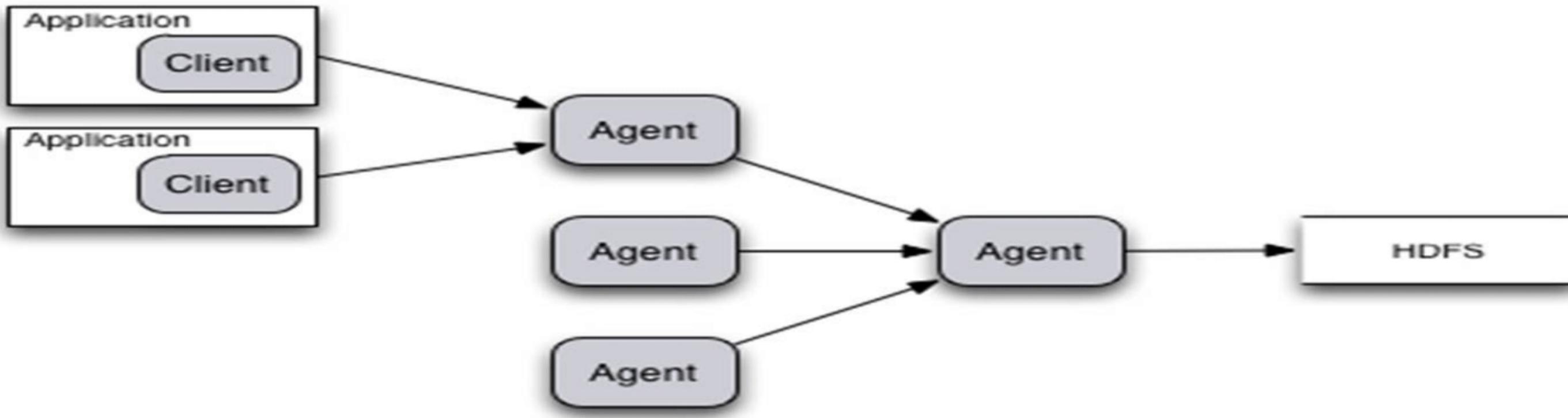
An entity that generates events and sends them to one or more Agents.

- Example
 - ❖ Flume log4j Appender
 - ❖ Custom Client using Client SDK (org.apache.flume.api)
- Decouples Flume from the system where event data is consumed from
- Not needed in all cases

A container for hosting Sources, Channels, Sinks and other components that enable the transportation of events from one place to another.

- Fundamental part of a Flume flow
- Provides Configuration, Life-Cycle Management, and Monitoring Support for hosted components

Typical Aggregation Flow



$[\text{Client}]^+ \rightarrow \text{Agent} [\rightarrow \text{Agent}]^* \rightarrow \text{Destination}$

An active component that receives events from a specialized location or mechanism and places it on one or Channels.

- Different Source types:
 - ❖ Specialized sources for integrating with well-known systems. Example: Syslog, Netcat
 - ❖ Auto-Generating Sources: Exec, SEQ
 - ❖ IPC sources for Agent-to-Agent communication: Avro
- Require at least one channel to function

- Reads data from the source system and passes onto the next hop or to the final destination.
- Flume Sources:
 - ❖ Avro Source
 - ❖ Exec Source
 - ❖ JMS Source
 - ❖ Spooling Directory Source

A passive component that buffers the incoming events until they are drained by Sinks.

- Different Channels offer different levels of persistence:
 - ❖ Memory Channel: volatile
 - ❖ File Channel: backed by WAL implementation
 - ❖ JDBC Channel: backed by embedded Database
- Channels are fully transactional
- Provide weak ordering guarantees
- Can work with any number of Sources and Sinks.

An active component that removes events from a Channel and transmits them to their next hop destination.

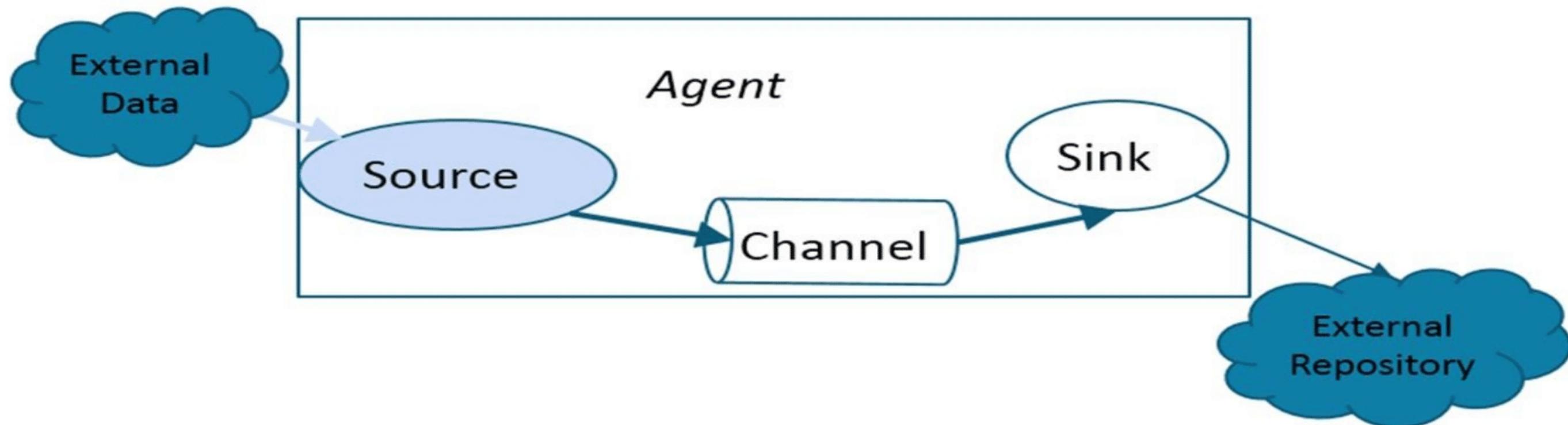
- Different types of Sinks:
 - ❖ Terminal sinks that deposit events to their final destination. For example: HDFS, HBase
 - ❖ IPC sink for Agent-to-Agent communication: Avro
- Require exactly one channel to function

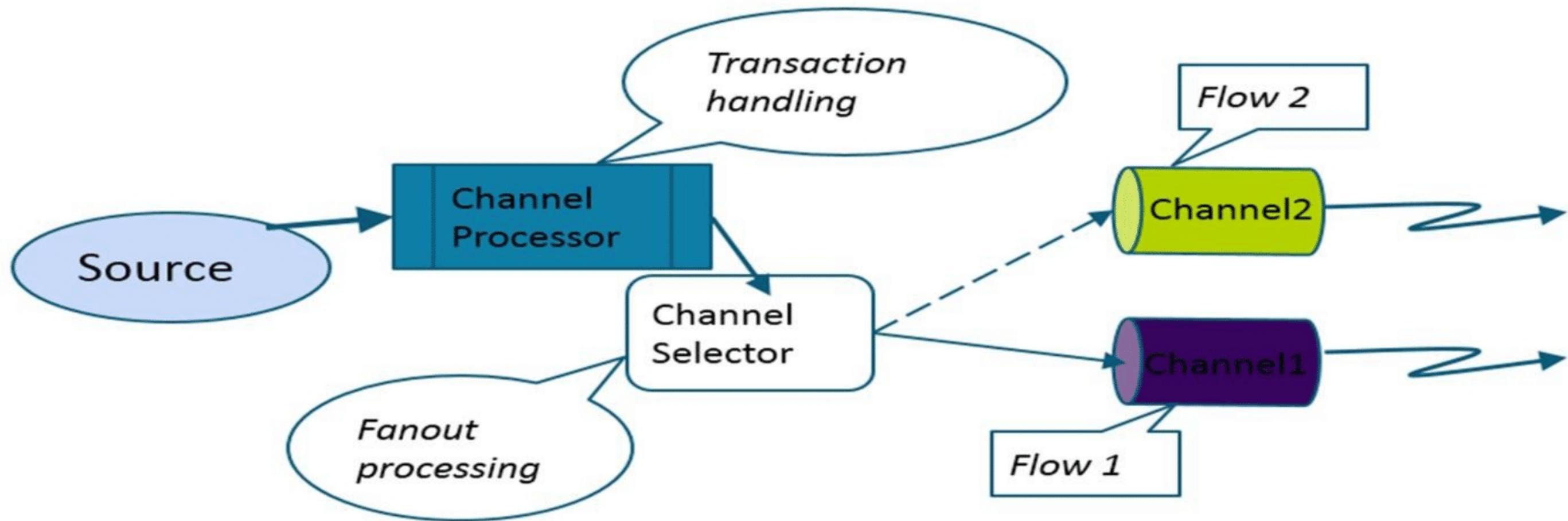
■ Writes data to the next hop or to the final destination.

■ Flume Sinks:

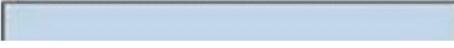
- ❖ Avro Sink
- ❖ HDFS Sink
- ❖ HBASE Sink
- ❖ File Sink
- ❖ Null Sink
- ❖ Logger Sink

What is the source in Flume





Current Flume Flow



FLUME ADOBE SYSTEMS

Monitoring: protocol support

- Several monitoring protocols supported out of the box
 - ❖ JMX
 - ❖ Ganglia
 - ❖ HTTP (JSON)

- Java opts must be set in flume-env.sh to configure monitoring
- Ganglia and HTTP monitoring are mutually exclusive

Thank You