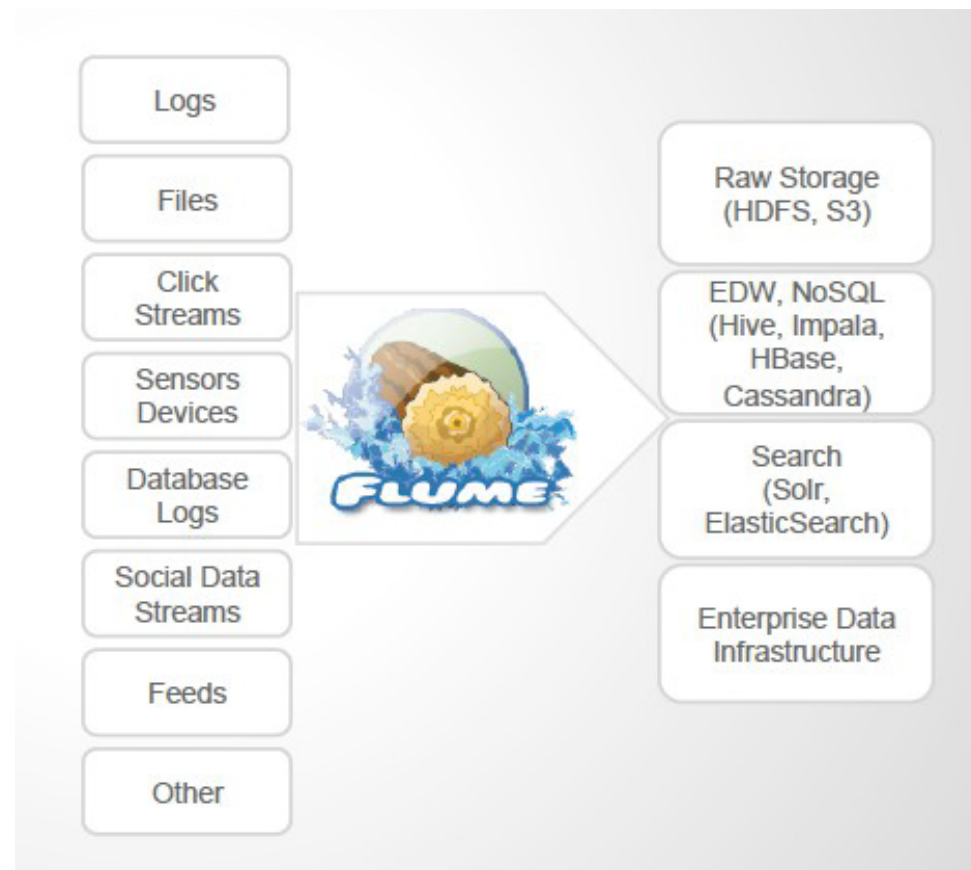# Apache Flume

Thaveewat Khanan

Introduction

Apache Flume is:

- **A distributed data transport and aggregation system for event- or log-structured data**
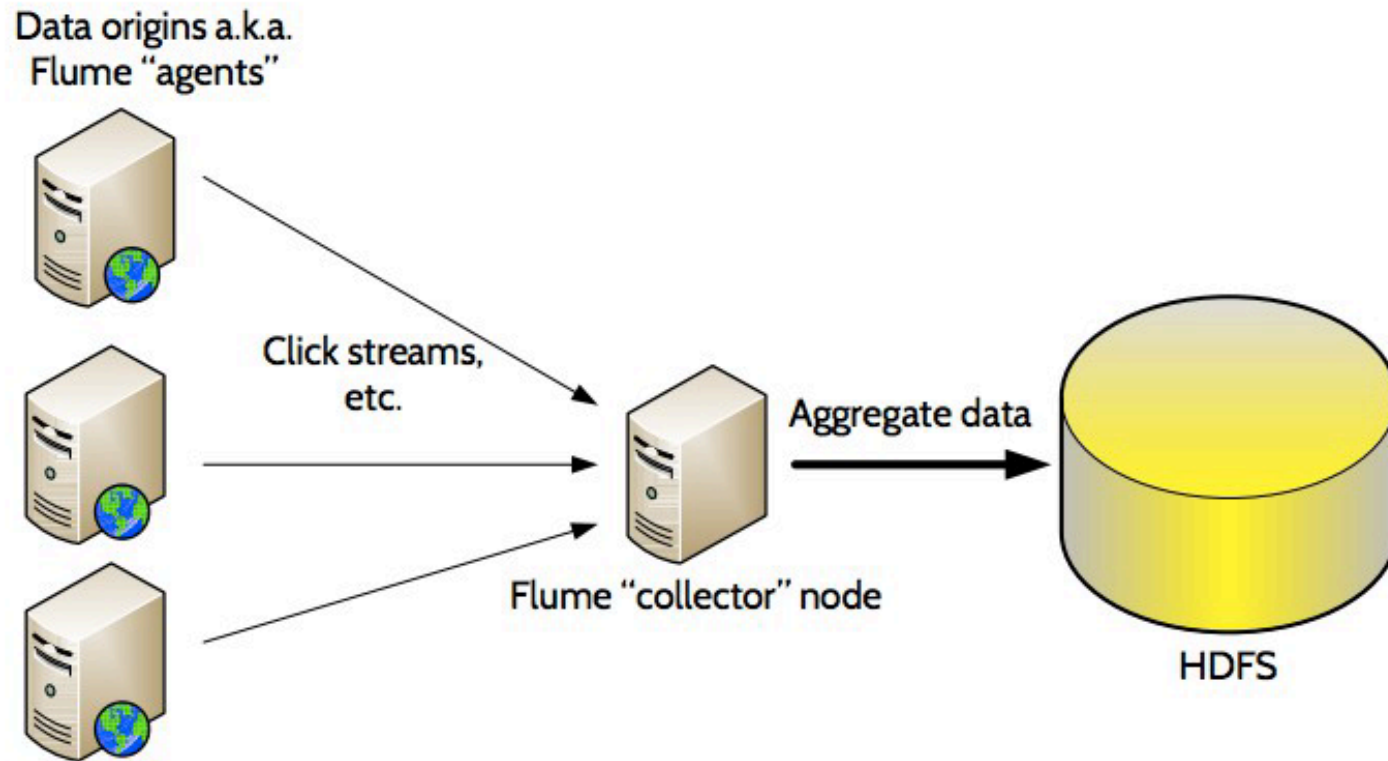- **Principally designed for continuous data ingestion into Hadoop... But more flexible than that**

What is Flume?

Apache Flume is a continuous data ingestion system that is...

- **open-source,**
- **reliable,**
- **scalable,**
- **manageable,**
- **Customizable,**
- **and designed for Big Data ecosystem**

Thaveewat Khanan

Architecture Overview

Data origins a.k.a. Flume "agents" → Click streams, etc. → Flume "collector" node → Aggregate data → HDFS
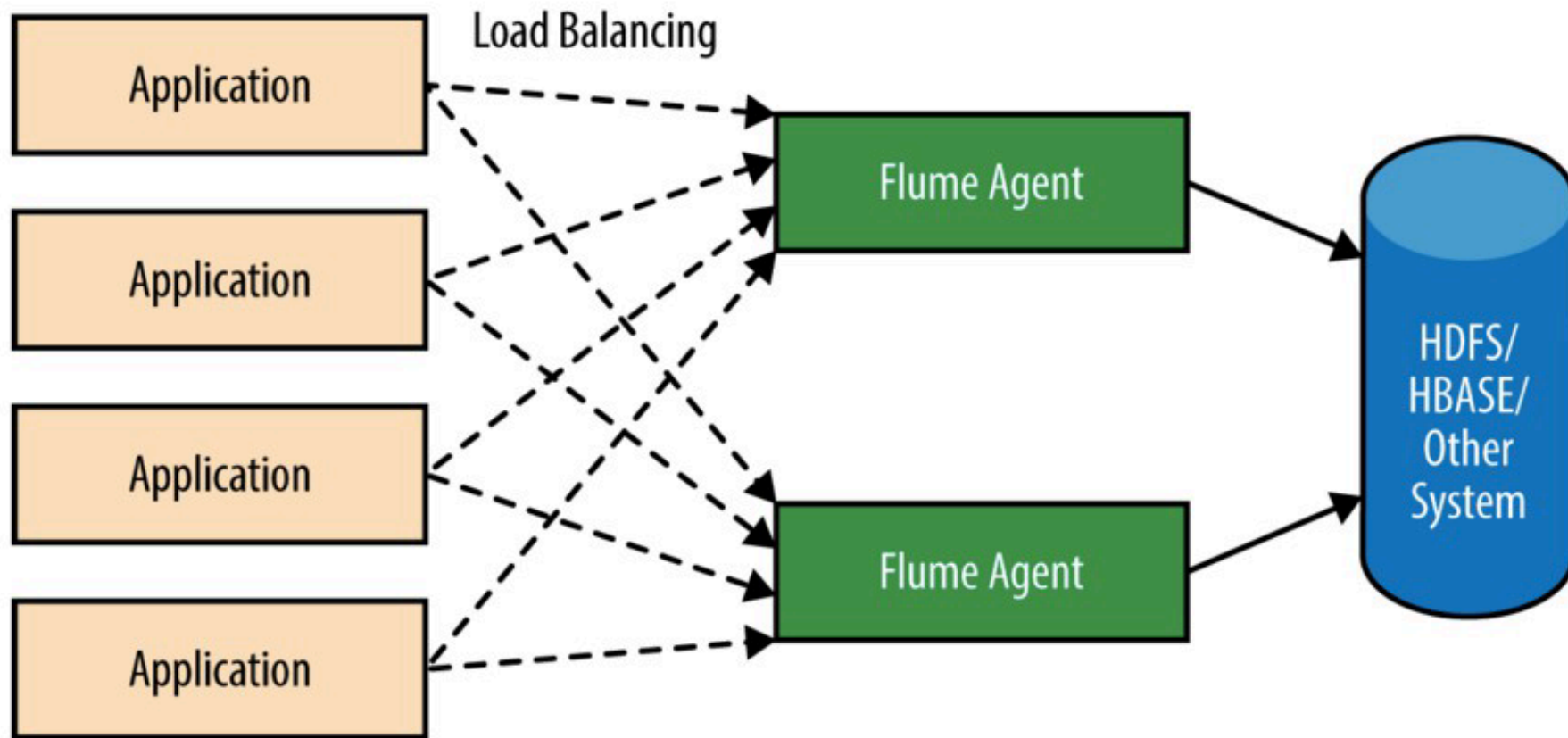
Thaveewat Khanan

Flume Agent

- **A source writes events to one or more channels.**

- **A channel is the holding area as events are passed from a source to a sink.**

- **A sink receives events from one channel only.**

- **An agent can have many channels.**

Thaveewat Khanan

Flow

Load Balancing

A Simple Flow

Source: Using Flume, Hari Shreedharan, 2014

Thaveewat Khanan

Flume Agent Configuration : Example

```
agent.sources = httpSrc
agent.channels = memory1 memory2
agent.sinks = hdfsSink hbaseSink

agent.sources.httpSrc.type = http
agent.sources.httpSrc.channels = memory1 memory2

# Bind to all interfaces
agent.sources.httpSrc.bind = 0.0.0.0
agent.sources.httpSrc.port = 4353

# Removing this line will disable SSL
agent.sources.httpSrc.ssl = true
agent.sources.httpSrc.keystore = /tmp/keystore
agent.sources.httpSrc.keystore-password = UsingFlume

agent.sources.httpSrc.handler = usingflume.ch03.HTTPSourceXMLHandler
agent.sources.httpSrc.handler.insertTimestamp = true

agent.sources.httpSrc.interceptors = hostInterceptor
agent.sources.httpSrc.interceptors.hostInterceptor.type = host
```

Thaveewat Khanan

Flume Agent Configuration : Example

```
# Initializes a memory channel with default configuration
agent.channels.memory1.type = memory

# Initializes a memory channel with default configuration
agent.channels.memory2.type = memory

# HDFS Sink
agent.sinks.hdfsSink.type = hdfs
agent.sinks.hdfsSink.channel = memory1
agent.sinks.hdfsSink.hdfs.path = /Data/UsingFlume/%{topic}/%Y/%m/%d/%H/%M
agent.sinks.hdfsSink.hdfs.filePrefix = UsingFlumeData

agent.sinks.hbaseSink.type = asynchbase
agent.sinks.hbaseSink.channel = memory2
agent.sinks.hbaseSink.serializer = usingflume.ch05.AsyncHBaseDirectSerializer
agent.sinks.hbaseSink.table = usingFlumeTable
```

Thaveewat Khanan

Stream Processing Architecture
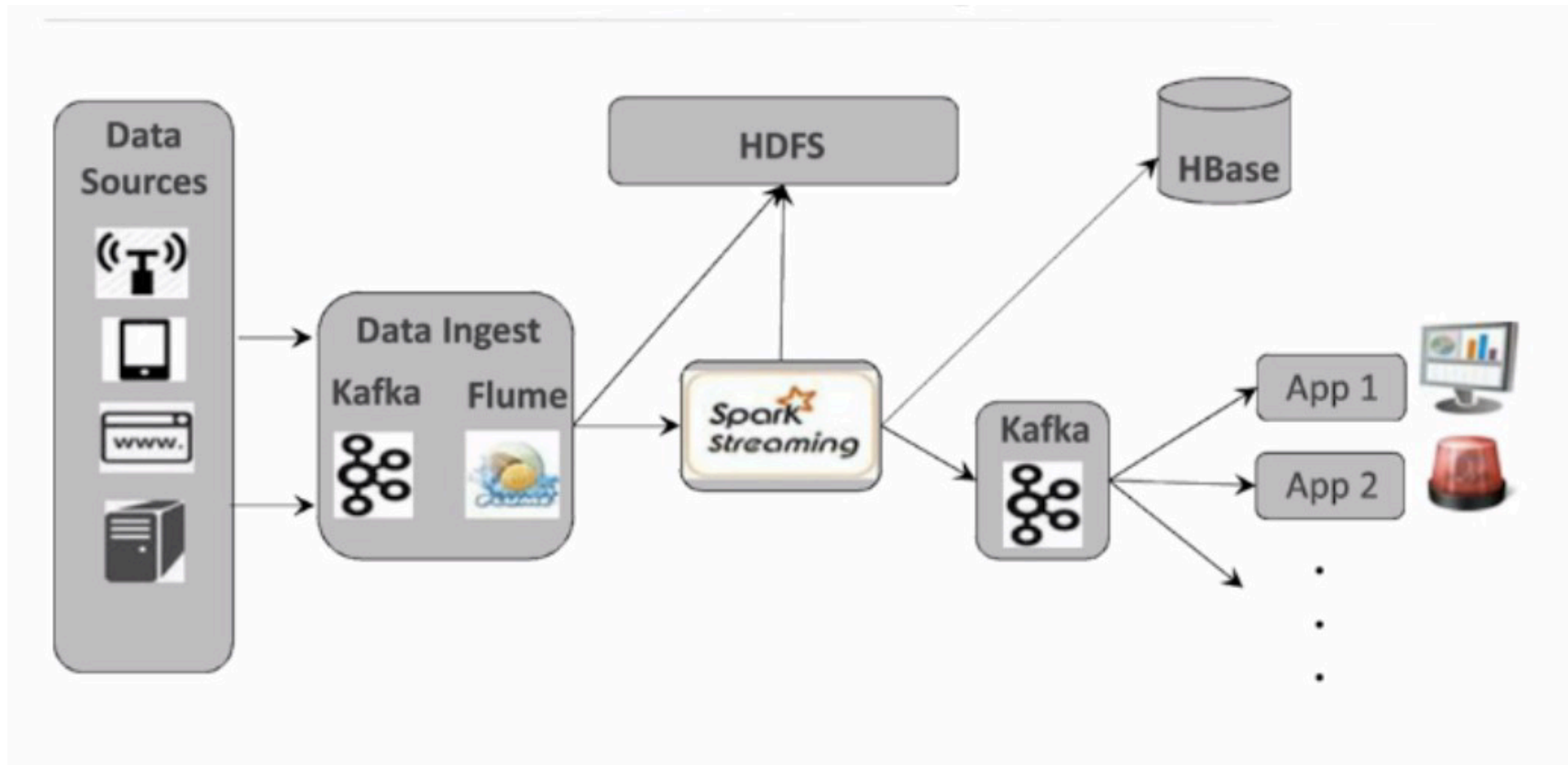
Flume Loading Data to HDFS

$ cd /etc/flume-ng/conf/

$ sudo rm flume.conf

$sudo wget https://github.com/bobbylovemovie/trainbigdata/raw/master/flume/flume.conf

$cat flume.conf

```
agent.sources = netsource
agent.sinks = hdfssink
agent.channels = memorychannel
agent.sources.netsource.type = netcat
agent.sources.netsource.bind = localhost
agent.sources.netsource.port = 3030
agent.sources.netsource.interceptors = ts
agent.sources.netsource.interceptors.ts.type = org.apache.flume.interceptor.TimestampInterceptor$Builder
agent.sinks.hdfssink.type = hdfs
agent.sinks.hdfssink.hdfs.path = hdfs://localhost:8020/user/cloudera/flume/events
agent.sinks.hdfssink.hdfs.filePrefix = log
agent.sinks.hdfssink.hdfs.rollInterval = 0
agent.sinks.hdfssink.hdfs.rollCount = 5
agent.sinks.hdfssink.hdfs.fileType = DataStream
agent.channels.memorychannel.type = memory
agent.channels.memorychannel.capacity = 100
agent.channels.memorychannel.transactionCapacity = 100
agent.sources.netsource.channels = memorychannel
agent.sinks.hdfssink.channel_= memorychannel
```

Thaveewat Khanan

Flume Loading Data to HDFS

start flume-service

$sudo service flume-ng-agent restart

```
[cloudera@quickstart ~]$ sudo service flume-ng-agent restart
Flume agent is not running                                    [  OK  ]
Starting Flume NG agent daemon (flume-ng-agent):              [  OK  ]
```

start flume Agent

$sudo flume-ng agent --conf /etc/flume-ng/conf/ --conf-file /etc/flume-ng/conf/flume.conf

--name agent -Dflume.root.logger=INFO,console

```
2016-10-31 03:20:39,476 (lifecycleSupervisor-1-3) [INFO - org.apache.flume.sourc
e.NetcatSource.start(NetcatSource.java:169)] Created serverSocket:sun.nio.ch.Ser
verSocketChannelImpl[/127.0.0.1:3030]
```

Thaveewat Khanan

Flume Loading Data to HDFS

Datasource Connect By Telnet

Open New Terminal

$sudo yum install telnet

$telnet localhost 3030

```
[cloudera@quickstart ~]$ telnet localhost 3030
Trying 127.0.0.1...
Connected to localhost.
Escape character is '^]'.
Hello Bigdata from Bobby Thaveewat
OK
a,b,c
OK
d,e,f
OK
g,h,i
OK
j,k,l
OK
```

Thaveewat Khanan

View Result

📄 File Browser

▦ View as binary

✏ Edit file

⬇ Download

📄 View file location

🔄 Refresh

Last modified
09/01/2017
4:01 AM

User
root

Group
cloudera

Size
28 B

🏠 Home

Page [1] to [1] of 1   ⏮ ⏪ ⏩ ⏭

/ user / cloudera / flume / events / **log.1504238424110**

```
a,b,c
d,e,f
g,h,i
j,k,l
```