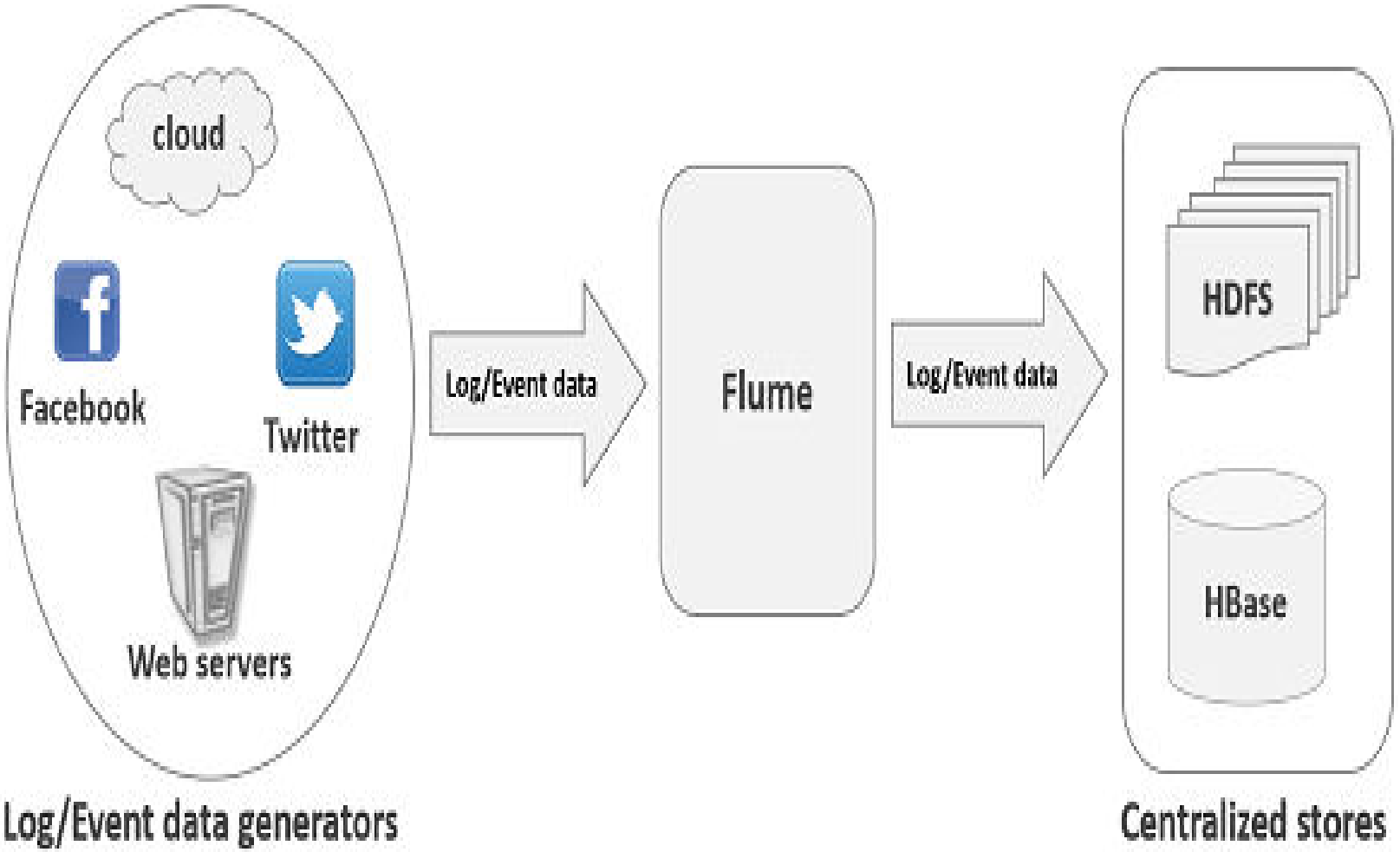
Apache Flume

What is Flume?

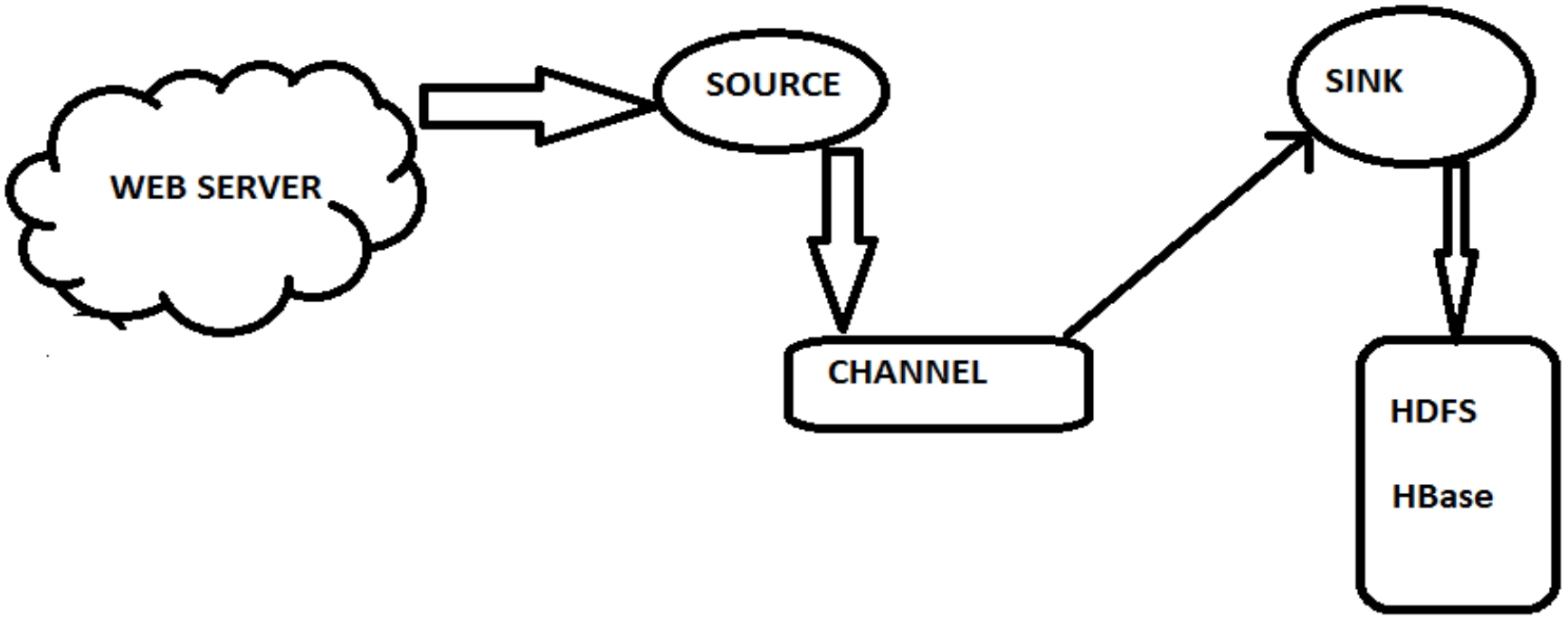
- **Apache Flume is a tool/service/data ingestion mechanism for collecting aggregating and transporting large amounts of streaming data such as log files, events (etc...) from various sources to a centralized data store.**

- **Flume is a highly reliable, distributed, and configurable tool. It is principally designed to copy streaming data (log data) from various web servers to HDFS.**

Flume Agent


cloud
Facebook
Twitter
Web servers
Log/Event data generators
Log/Event data
Flume
Log/Event data
HDFS
HBase
Centralized stores

Flume Components



Advantages of Flume

- **Using Apache Flume we can store the data in to any of the centralized stores (HBase, HDFS).**
- **When the rate of incoming data exceeds the rate at which data can be written to the destination, Flume acts as a mediator between data producers and the centralized stores and provides a steady flow of data between them.**
- **The transactions in Flume are channel-based where two transactions (one sender and one receiver) are maintained for each message. It guarantees reliable message delivery.**
- **Flume is reliable, fault tolerant, scalable, manageable, and customizable.**

Features of Flume

- **Flume ingests log data from multiple web servers into a centralized store (HDFS, HBase) efficiently.**
- **Using Flume, we can get the data from multiple servers immediately into Hadoop.**
- **Along with the log files, Flume is also used to import huge volumes of event data produced by social networking sites like Facebook and Twitter, and e-commerce websites like Amazon and Flipkart.**
- **Flume supports a large set of sources and destinations types.**

Flume Event

- An **event** is the basic unit of the data transported inside **Flume**. It contains a payload of byte array that is to be transported from the source to the destination accompanied by optional headers.

Flume Component

- **Source**
- **A source is the component of an Agent which receives data from the data generators and transfers it to one or more channels in the form of Flume events.**
- **Apache Flume supports several types of sources and each source receives events from a specified data generator.**

Channel

- A channel is a transient store which receives the events from the source and buffers them till they are consumed by sinks. It acts as a bridge between the sources and the sinks.
- These channels are fully transactional and they can work with any number of sources and sinks.
- Example – JDBC channel, File system channel, Memory channel, etc.

Sink

- A sink stores the data into centralized stores like HBase and HDFS. It consumes the data (events) from the channels and delivers it to the destination. The destination of the sink might be another agent or the central stores.
- Example – HDFS sink

Basic Rules

- Every agent must have at least one channel
- Every source must have at least one channel.
- Every sink must have exactly one channel.
- Every component must have a type.