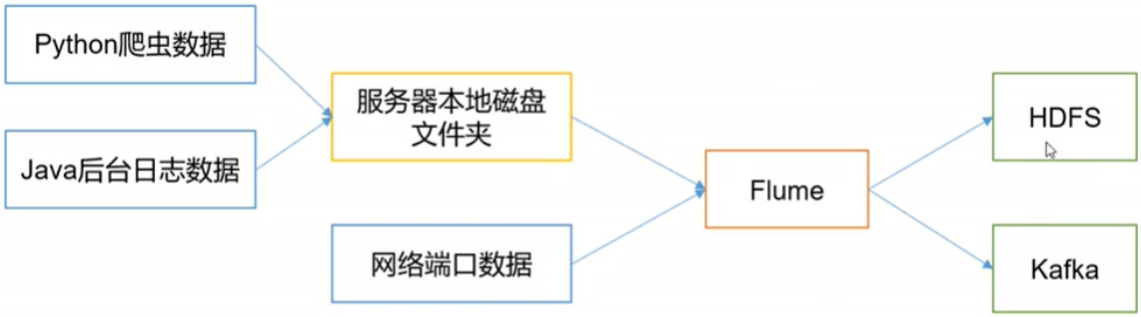
Flume1 概述

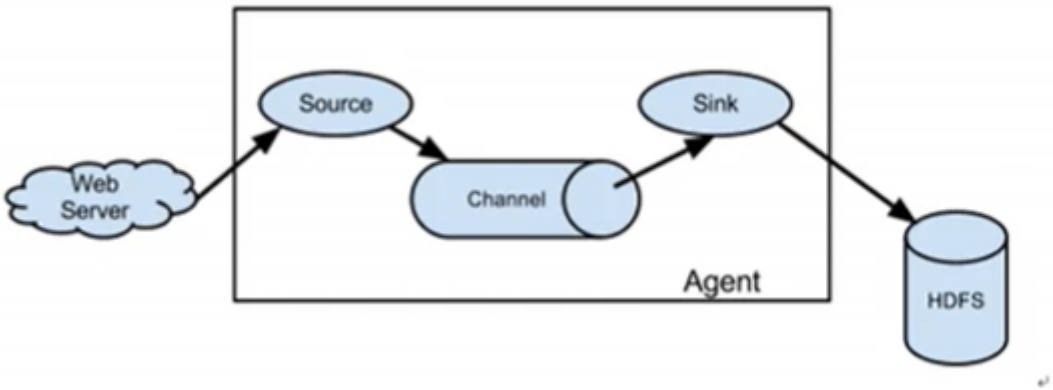
定义

Flume是Cloudera提供一个高可用的，高可靠的，分布式的海量日志采集、聚合和传输的系统，Flume基于流失架构，灵活简单。



Flume主要用于对Java后台日志以及Python爬虫数据的数据实时读取进行一个传输，传输到HDFS中。

基础架构



1. source: 一个存储数据的地方，将读入的数据存储起来
2. channel：管道，将数据运输到输出端口的管道(防止读入比写出快而导致崩溃)
3. sink: 输出端口，常用于连接HDFS或者Kafka

组件介绍

Agent

Agent是一个JVM进程，它以事件的形式将数据从源头送至目的
Agent主要有3个部分组成： Source、Channel、Sink

Source

Source是负责接收数据到 Flume Agent的组件，Source组件可以处理各种类型、各种格式的日志数据，包括Avro、Thrift、Exec、Jms、Spooling、Netcat、Sequence、Generator、Syslog、Http、Legacy

Sink

Sink不断地轮询Channel中的事件且批量移除它们，并将这些事件批量写入到存储或者索引系统、或者被发送到另一个Flume Agent。
Sinl组件目的地包括 HDFS、 Logger、Avro、Thrift、IPC、File、HBase、slor、自定义

Channel

channel是位于Source和Sink之间的缓冲区。因此，Channel允许Source和Sink运用再不同速率上。Channel是线程安全的，可以同时处理几个Source的写入操作和几个Sink的读取操作。

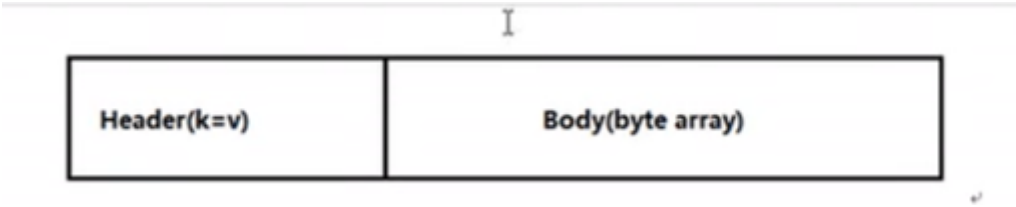
Flume 自带两种 Channel： Memory Channel 和 File Channel 以及 KafakaChannel

Memory Channel 是内存中的队列。Memory Channel 在不需要关心数据丢失的情景下适用。如果需要关心数据丢失，那么Memory Channel 就不应该使用，因为程序死亡，机器宕机或者重启都会导致数据丢失。

File Channel 将所有事件写到磁盘，因此在程序关闭或机器宕机的情况下不会丢失数据

Event

传输单元，Flume 数据传输的基本单元，以Event的形式将数据从源头送至目的地。Event由Header和Body两部分组成，Header用来存放event的一些属性，为K-V结构，Boby用来存放该条数据，形式为字节数据。



常用的 Flume Source的组件有

1. Avro Source
2. Exec Source
3. Taildir Source
4. Kafka Source
5. NetCat Source

常用的 Flume Sinks的组件有

```
1. HDFS Sink
2. Hive Sink
3. Logger Sink
4. Avro Sink
5. File Roll Sink
6. Kafka Sink
```

常用的Flume Channels的组件有

```
1. Memory Channel
2. JDBC Channel
3. Kafka Channel
4. File Channel
```