课程大纲（辅助系统）

前言

在一个完整的离线大数据处理系统中，除了 hdfs+mapreduce+hive 组成分析系统的核心之外，还需要数据采集、结果数据导出、任务调度等不可或缺的辅助系统，而这些辅助工具在 hadoop 生态体系中都有便捷的开源框架：

日志采集框架 Flume

Flume 介绍

概述

- Flume 是一个分布式、可靠、和高可用的海量日志采集、聚合和传输的系统。
- Flume 可以采集文件，socket 数据包、文件、文件夹、kafka 等各种形式源数据，又可以将采集到的数据(下沉 sink)输出到 HDFS、hbase、hive、kafka 等众多外部存储系统中
- 一般的采集需求，通过对 flume 的简单配置即可实现
- Flume 针对特殊场景也具备良好的自定义扩展能力，

因此，flume 可以适用于大部分的日常数据采集场景

运行机制

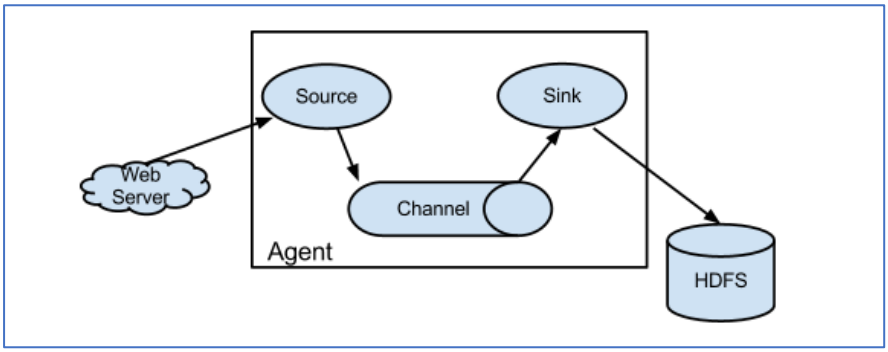
1、 Flume 分布式系统中最**核心的角色是 agent**，flume 采集系统就是由一个个 agent 所连接起来形成

2、 每一个 agent 相当于一个数据传递员，内部有三个组件：

a) Source：采集组件，用于跟数据源对接，以获取数据

b) Sink：下沉组件，用于往下一级 agent 传递数据或者往最终存储系统传递数据

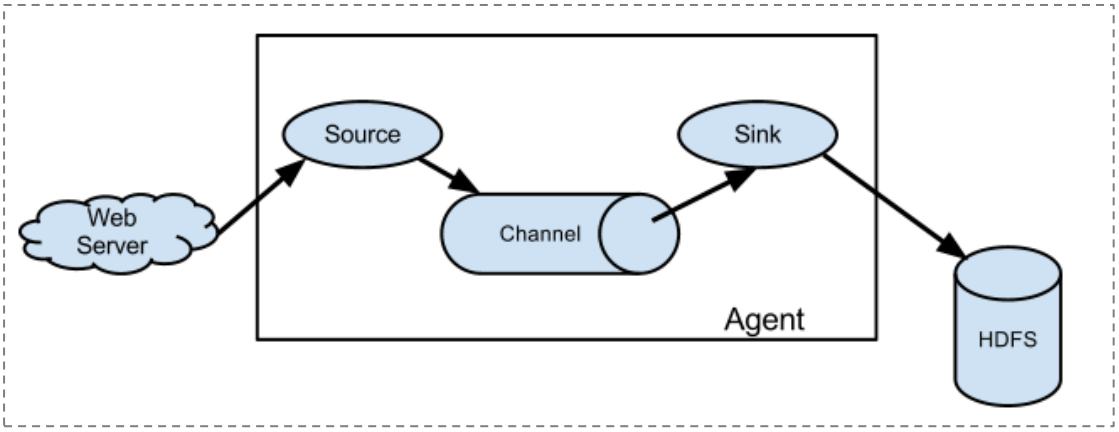
c) Channel：传输通道组件，用于从 source 将数据传递到 sink


Source
Sink
Web
Server
Channel
Agent
HDFS

Flume 采集系统结构图

1. 简单结构

单个 agent 采集数据



2. 复杂结构

多级 agent 之间串联