第一步：

用代码模拟客户端上传日志信息，埋点信息

第二步：

设置 nginx 配置信息，收集埋点服务

第三步：

进行 flume 的配置，采集 nginx 的生成的.log 文件的日志信息。

并进行写入到 HDFS 中。

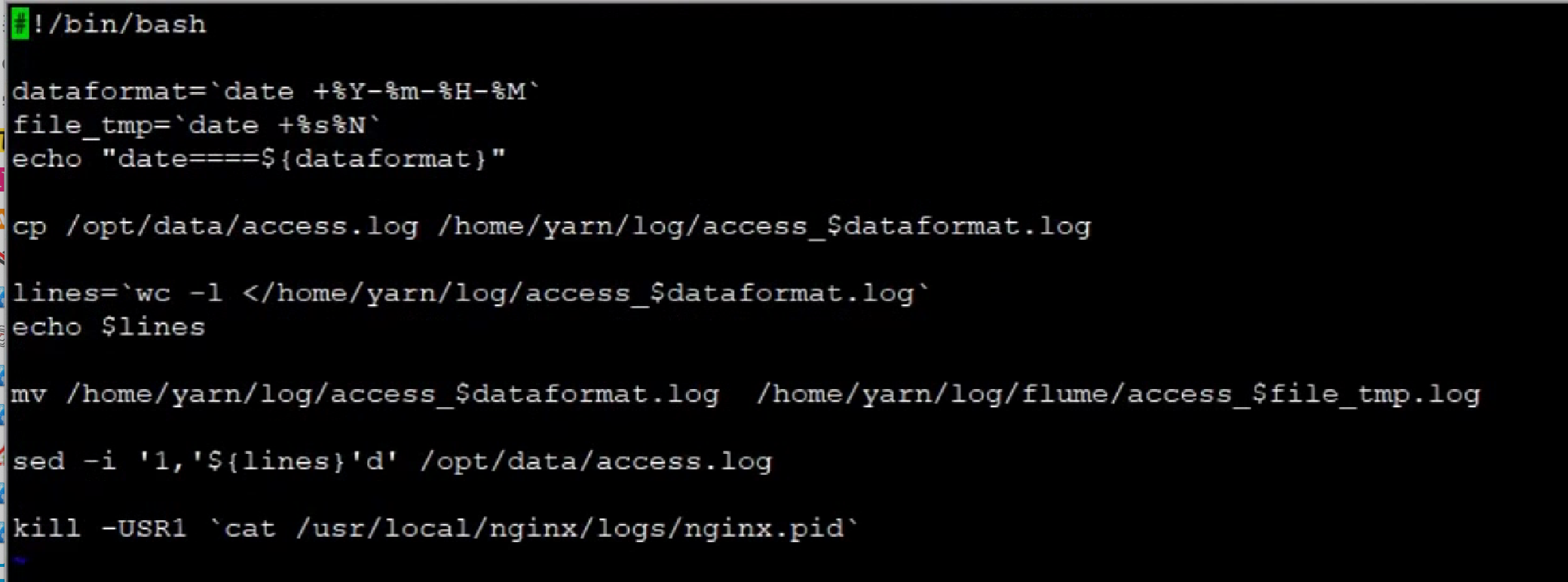
flume：

支持多输入到一个输出，一个输入到多个输出

flume 学习播客：https://www.cnblogs.com/zhangyinhua

nginx 埋点日志

滚动日志文件脚本内容：



第11讲 flume几种常见拦截器介绍以及正则过滤文件数据

到这里的时候，就会有个问题，比如说 server\_time字段，默认的是毫秒单位的时间戳，

但是写到的目录是2018-11-16 的单位目录的。

>> server\_time = 1542069815000

>> 落地 HDFS = 2018-11-16

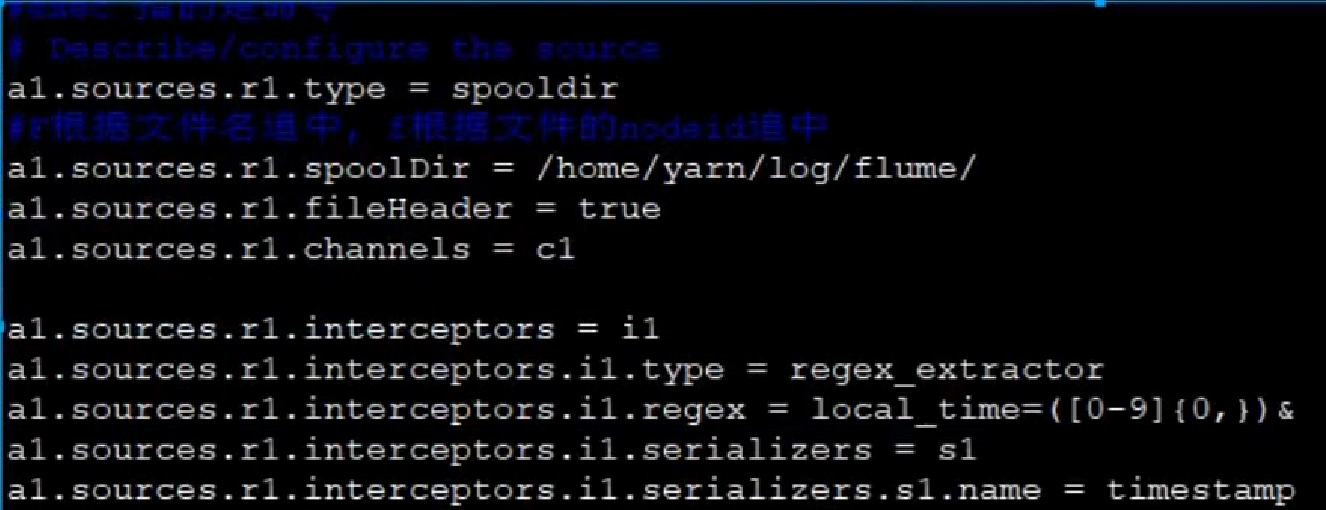
这里讲解一下 flume 的拦截器，

参考资料地址：http://www.lxw1234.com/archives/2015/11/543.htm

这里还涉及到自定义 tailsource 的讲解，可以自己谷歌了。

上面的链接中介绍了很多的过滤器，这里我们选择用Regex Filter Interceptor正则过滤器来解决掉问题

第12讲 flume解析行为日志的日期把数据插入对应的hdfs分区



第13讲 hive和spark介绍以及2者区别