流式数据解决方案框架一

适用应用场景

流式数据

低延时

硬件环境

至少3台服务器。

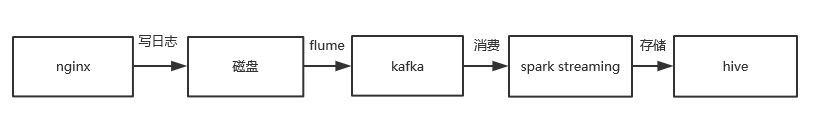
软件环境

CentOs 7.0

Jdk1.7/jdk1.8

...

架构



处理流程

Flume会[实时监控](https://www.baidu.com/s?wd=%E5%AE%9E%E6%97%B6%E7%9B%91%E6%8E%A7&tn=24004469_oem_dg&rsv_dl=gh_pl_sl_csd" \t "https://blog.csdn.net/trigl/article/details/_blank)写入日志的磁盘，只要有新的日志写入，Flume就会将日志以消息的形式传递给Kafka，然后Spark Streaming实时消费消息传入Hive

预期效果

通过这套架构，收集到的日志可以及时被Flume发现传到Kafka，通过Kafka我们可以把日志用到各个地方，同一份日志可以存入Hdfs中，也可以离线进行分析，还可以实时计算，而且可以保证安全性，基本可以达到实时的要求。

效果计算分析

每秒多少条，一天处理多少数据量，计算时效性

部署流程

1. 安装、配置flume
2. 安装、配置kafka
3. 安装、配置zookeeper
4. 安装、配置hadoop
5. 安装、配置hive
6. 安装、配置spark
7. 配置开发环境

开发