# Step1

## 数据采集

通过页面嵌入JS代码的方式获取用户访问行为，并发送到web服务的后台记录日志

然后，将各服务器上生成的点击流日志通过实时或批量的方式汇聚到HDFS文件系统中

上传数据到hdfs

hadoop fs –put xxx.txt /data

# Step2 数据预处理

通过mapreduce程序对采集到的点击流数据进行预处理，比如清洗，格式整理，滤除脏数据等

### 原始数据的一条字段解析

*58.215.204.118 - - [18/Sep/2013:06:51:35 +0000] "GET /wp-includes/js/jquery/jquery.js?ver=1.10.2 HTTP/1.1" 304 0 "http://blog.fens.me/nodejs-socketio-chat/" "Mozilla/5.0 (Windows NT 5.1; rv:23.0) Gecko/20100101 Firefox/23.0"*

1.访客ip地址：*58.215.204.118*

*2.访客用户信息：- -*

*3.访问时间：[18/Sep/2013:06:51:35 +0000]*

*4.请求方式：GET*

*5.请求url：/wp-includes/js/jquery/jquery.js?ver=1.10.2*

6.请求协议：*HTTP/1.1"*

*7.响应吗：304*

*8.返回的访问数据流量*

*9.访客来源url：*[*http://blog.fens.me/nodejs-socketio-chat/*](http://blog.fens.me/nodejs-socketio-chat/)

*10.访客所用浏览器："Mozilla/5.0 (Windows NT 5.1; rv:23.0) Gecko/20100101 Firefox/23.0"*

### 编写MR程序进行预处理

见代码

### 运行MR程序

1.将程序打成JAR包，放到HDFS，然后在集群的任意一个节点上用hadoop命令启动

hadoop fs –put d:/ WeblogPreProcess.jar /weblog/

2.把源数据放到hdfs上面

hadoop fs –put weblog.txt /weblog/input

3.运行jar包

hadoop jar WeblogPreProcess.jar com.lzy.main.WeblogPreprocess /weblog/input /weblog/output

# Step3 数据入库

将预处理之后的数据导入到HIVE仓库中相应的库和表中

# Step4 数据分析

项目的核心内容，即根据需求开发ETL分析语句，得出各种统计结果

# Step5 数据展现

将分析所得数据进行可视化