# 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

flume采集数据

# 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

### 根据清洗后数据梳理出pageviews模型数据

1. 用于生成点击流的访问日志表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 时间戳 | IP地址 | Cookie | Session | 请求URL | Referal |
| 2012-01-01 12:31:12 | 101.0.0.1 | User01 | S001 | /a/... | somesite.com |
| 2012-01-01 12:31:16 | 201.0.0.2 | User02 | S002 | /a/... | - |
| 2012-01-01 12:33:06 | 101.0.0.2 | User03 | S002 | /b/... | baidu.com |
| 2012-01-01 15:16:39 | 234.0.0.3 | User01 | S003 | /c/... | google.com |
| 2012-01-01 15:17:11 | 101.0.0.1 | User01 | S004 | /d/... | /c/... |
| 2012-01-01 15:19:23 | 101.0.0.1 | User01 | S004 | /e/... | /d/.... |

1. 页面点击流模型Pageviews表(按session聚集的访问页面信息)(每个session中的每个url也即是访问页面,的记录信息, 想差半个小时了就认为是下一个session了)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Session | userid | 时间 | 访问页面URL | 停留时长 | 第几步 |
| S001 | User01 | 2012-01-01 12:31:12 | /a/.... | 30 | 1 |
| S002 | User02 | 2012-01-01 12:31:16 | /a/.... | 10 | 1 |
| S002 | User02 | 2012-01-01 12:33:06 | /b/.... | 110 | 2 |
| S002 | User02 | 2012-01-01 12:35:06 | /e/.... | 30 | 3 |

1. 点击流模型Visits表

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Session | 起始时间 | 结束时间 | 进入页面 | 离开页面 | 访问页面数 | IP | cookie | referal |
| S001 | 2012-01-01 12:31:12 | 2012-01-01 12:31:12 | /a/... | /a/... | 1 | 101.0.0.1 | User01 | somesite.com |
| S002 | 2012-01-01 12:31:16 | 2012-01-01 12:35:06 | /a/... | /e/... | 3 | 201.0.0.2 | User02 | - |
| S003 | 2012-01-01 12:35:42 | 2012-01-01 12:35:42 | /c/... | /c/... | 1 | 234.0.0.3 | User03 | baidu.com |
| S004 | 2012-01-01 15:16:39 | 2012-01-01 15:19:23 | /c/... | /e/... | 3 | 101.0.0.1 | User01 | google.com |
| …… | …… | …… | …… | …… | …… | …… | …… | …… |

这就是点击流模型。当WEB日志转化成点击流数据的时候，很多网站分析度量的计算变得简单了，这就是点击流的“魔力”所在。基于点击流数据我们可以统计出许多常见的网站分析度量

# 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

## 创建贴源数据表

drop table if exists ods\_weblog\_origin;

create table ods\_weblog\_origin(

valid string,

remote\_ip string,

remote\_user string,

time\_local string,

request string,

status string,

body\_bytes\_sent string,

http\_referer string,

http\_user\_agent string

)

partitioned by (datestr string)

row format delimited fields terminated by '\001';

导入清洗结果数据到贴源数据表ods\_weblog\_origin

load data inpath '/weblog/preprocessed/16-02-24-16/' overwrite into table ods\_weblog\_origin partition(datestr='2013-09-18');

## 创建点击流模型表

drop table if not exists ods\_click\_pageviews;

create table ods\_click\_pageviews(

session string,

user\_ip string,

visit\_date string,

request\_url string,

stay\_time string,

visit\_step

)

partitioned by (datestr string)

row format delimited fields terminated by ‘\001’

导入点击流模型pageviews数据到ods\_click\_pageviews表

load data inpath '/weblog/clickstream/pageviews' overwrite into table ods\_click\_pageviews partition(datestr='2013-09-18');

# 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

# Hive的安装

# Flume配置