

Hive综合

高校大数据课程系列

ENTER

课程目标

Course objectives

掌握Hive综合
应用

1

掌握受访与访客
分析案例

3

掌握流量统计
分析案例

2

本章任务

Task of this chapter

1

流量统计分析

2

受访与访客分析

任务1 流量统计分析

任务背景

网站是由多个网页（Page）构成，当用户在访问多个网页时，网页与网页之间是靠Referrers参数来标识上级网页来源。由此，可以确定网页被依次访问的顺序，当然也可以通过时间来标识访问的次序。其次，用户对网站的每次访问，可视为是一次会话（Session），在网站日志中将会用不同的Sessionid来唯一标识每次会话。如果把Page视为“点”的话，那么我们可以很容易的把Session描绘成一条“线”，也就是用户的点击流数据轨迹曲线。对于网页之间的分析我们可以通过访问web日志来完成。可以分析出页面访问的PV, UV，人均访问次数。页面访问频率等。对于统计各业务指标有着基础的支撑作用。

任务1 流量统计分析

任务需求

利用数据仓库Hive完成Web日志的分析。

1. web数据集1的格式如下，各列的含义分别是：是否合法、来源的IP地址、客户端的用户名、访问时间与时区、请求的URL、请求的状态码、发送客户端的字节数、客户端浏览器的相关信息。

```
false||194.237.142.21||-||2013-09-18 06:49:18||/wp-content/uploads/2013/07/rstudio-git3.png||304||0||"-||"Mozilla/4.0(compatible;)"
```

```
false||163.177.71.12||-||2013-09-18 06:49:33||/||200||20||"-||"DNSPod-Monitor/1.0"
```

```
false||163.177.71.12||-||2013-09-18 06:49:36||/||200||20||"-||"DNSPod-Monitor/1.0"
```

```
false||101.226.68.137||-||2013-09-18 06:49:42||/||200||20||"-||"DNSPod-Monitor/1.0"
```

```
false||101.226.68.137||-||2013-09-18 06:49:45||/||200||20||"-||"DNSPod-Monitor/1.0"
```

参考

Hive综合实验(流量统计分析)-实验手册

任务1 流量统计分析

任务需求

利用数据仓库Hive完成Web日志的分析。

2. web数据集2的格式如下，各列的含义分别是：会话SessionID、来源的IP地址、客户端的用户名、访问时间与时区、请求的URL、访问的步骤、页面停留时长、来源URL、用户客户端信息、发送的字节数、状态码。

```
99b1210d-c77a-4ac0-a1de-1c2db65b31c7||1.80.249.223||-||2013-09-18 07:57:33|/hadoop-hive-  
intro/||1||60||"http://www.google.com.hk/url?sa=t&rct=j&q=hive%E7%9A%84%E5%AE%89%E8%A3%85&source=web&cd=  
2&ved=0CC4QFjAB&url=%68%74%74%70%3a%2f%2f%62%6c%6f%67%2e%66%65%6e%73%2e%6d%65%2f%68%61%64%6f%6f%70%2d%68  
%69%76%65%2d%69%6e%74%72%6f%2f&ei=5lw5Uo-  
2NpGZiQfCwoG4BA&usg=AFQjCNF8EFxPuCMrm7CvqVgzcBUzrJZStQ&bvm=bv.52164340,d.aGc&cad=rjt"||"Mozilla/5.0(Wind  
owsNT5.2;rv:23.0)Gecko/20100101Firefox/23.0"||14764||20
```

参考

Hive综合实验(流量统计
分析)-实验手册

任务1 流量统计分析

任务需求

利用数据仓库Hive完成Web日志的分析。

3. web数据集3的格式如下，各列的含义分别是：会话SessionID、来源的IP地址、访问开始时间、访问结束时间、访问开始页面、访问结束页面、来源URL、访问的页面数。

```
0833aba0-498d-4758-80d9-6ac4414ddf8e||123.116.73.157||2013-09-19 00:58:58||2013-09-19 00:58:58||/hadoop-zookeeper-intro/||/hadoop-zookeeper-intro/||"https://www.google.com.hk/"||1
```

```
138a8025-730a-47ad-a376-4e752e5fe5cb||174.120.8.226||2013-09-18 13:22:30||2013-09-18 13:27:03||/hadoop-mahout-roadmap/||/hadoop-mahout-roadmap/||"-||4
```

参考

Hive综合实验(流量统计分析)-实验手册

任务1 流量统计分析

任务需求

利用数据仓库Hive完成Web日志的分析。

4. 数据集4的格式如下，各列的含义分别是：序号，年、月、日、小时。

1, 2013, 09, 17, 22

2, 2013, 09, 17, 23

3, 2013, 09, 18, 00

4, 2013, 09, 18, 01

5, 2013, 09, 18, 02

6, 2013, 09, 18, 03

参考

7-1. Hive综合实验(流量统计分析)-实验手册

任务1 流量统计分析

任务需求

利用数据仓库Hive完成Web日志的分析。

5. 完成以下数据处理业务。

- 1). 创建数据仓库表
- 2). 数据导入表
- 3). 创建明细宽表
- 4). 计算每小时pvs
- 5). 计算（一天）中的各小时pvs
- 6). 计算每天的pvs
- 7). 计算每月的pvs
- 8). 统计每小时各来访url产生的pv量
- 9). 统计每小时各来访host的产生的pv数并排序
- 10). 统计一天内各小时产生最多pvs的来源topN
- 11). 统计20130918日所有来访者平均请求的页面数。

参考

Hive综合实验(流量
统计分析)-实验手
册

任务1 流量统计分析

任务分析

首先创建ODS层表，ODS（Operational Data Store）是数据仓库体系结构中的一个可选部分，也被称为贴源层。ODS具备数据仓库的部分特征和OLTP系统的部分特征，它是“面向主题的、集成的、当前或接近当前的、不断变化的”数据。把数据再导入到ODS层表，再创建ODS层宽表用于统计分析。中间阶段可以创建中间表最后汇总到DW中。

任务1 流量统计分析

任务步骤

- 1、启动Hadoop伪分布运行环境
- 2、启动Hive
- 3、创建数据仓库及表
- 4、导入数据
- 5、数据分析

任务1 流量统计分析

任务结果

根据需求执行SQL语句，展示每个SQL的数据结果。

本章任务

Task of this chapter

1

流量统计分析

2

受访与访客分析

任务2 受访与访客分析

任务背景

网站是由多个网页（Page）构成，当用户在访问多个网页时，网页与网页之间是靠Referrers参数来标识上级网页来源。由此，可以确定网页被依次访问的顺序，当然也可以通过时间来标识访问的次序。其次，用户对网站的每次访问，可视为是一次会话（Session），在网站日志中将会用不同的Sessionid来唯一标识每次会话。如果把Page视为“点”的话，那么我们可以很容易的把Session描绘成一条“线”，也就是用户的点击流数据轨迹曲线。对于网页之间的分析我们可以通过访问web日志来完成。可以分析出页面访问的PV, UV，人均访问次数。页面访问频率等。对于统计各业务指标有着基础的支撑作用。

任务2 受访与访客分析

任务需求

利用数据仓库Hive完成Web日志的分析。

1. web数据集1的格式如下，各列的含义分别是：是否合法、来源的IP地址、客户端的用户名、访问时间与时区、请求的URL、请求的状态码、发送客户端的字节数、客户端浏览器的相关信息。

```
false||194.237.142.21||-||2013-09-18 06:49:18||/wp-content/uploads/2013/07/rstudio-  
git3.png||304||0||"-||"Mozilla/4.0(compatible;)"
```

```
false||163.177.71.12||-||2013-09-18 06:49:33||/||200||20||"-||"DNSPod-Monitor/1.0"
```

```
false||163.177.71.12||-||2013-09-18 06:49:36||/||200||20||"-||"DNSPod-Monitor/1.0"
```

```
false||101.226.68.137||-||2013-09-18 06:49:42||/||200||20||"-||"DNSPod-Monitor/1.0"
```

```
false||101.226.68.137||-||2013-09-18 06:49:45||/||200||20||"-||"DNSPod-Monitor/1.0"
```

参考

Hive综合实验(受访与访客分析)-实验手册

任务2 受访与访客分析

任务需求

利用数据仓库Hive完成Web日志的分析。

2. web数据集2的格式如下，各列的含义分别是：会话SessionID、来源的IP地址、客户端的用户名、访问时间与时区、请求的URL、访问的步骤、页面停留时长、来源URL、用户客户端信息、发送的字节数、状态码。

```
99b1210d-c77a-4ac0-a1de-1c2db65b31c7||1.80.249.223||-||2013-09-18 07:57:33||/hadoop-hive-  
intro/||1||60||"http://www.google.com.hk/url?sa=t&rct=j&q=hive%E7%9A%84%E5%AE%89%E8%A3%85&source  
=web&cd=2&ved=0CC4QFjAB&url=%68%74%74%70%3a%2f%2f%62%6c%6f%67%2e%66%65%6e%73%2e%6d%65%2f%68%61%6  
4%6f%6f%70%2d%68%69%76%65%2d%69%6e%74%72%6f%2f&ei=5lw5Uo-  
2NpGZiQfCwoG4BA&usg=AFQjCNF8EFxPuCMrm7CvqVgzCBUzrJZStQ&bvm=bv.52164340,d.aGc&cad=rjt"||"Mozilla/  
5.0 (Windows NT 5.2; rv:23.0) Gecko/20100101 Firefox/23.0"||14764 ||20
```

参考

Hive综合实验(受访
与访客分析)-实验
手册

任务2 受访与访客分析

任务需求

利用数据仓库Hive完成Web日志的分析。

3. web数据集3的格式如下，各列的含义分别是：会话SessionID、来源的IP地址、访问开始时间、访问结束时间、访问开始页面、访问结束页面、来源URL、访问的页面数。

```
0833aba0-498d-4758-80d9-6ac4414ddf8e||123.116.73.157||2013-09-19 00:58:58||2013-09-19  
00:58:58||/hadoop-zookeeper-intro/||/hadoop-zookeeper-intro/||"https://www.google.com.hk/"||1  
  
138a8025-730a-47ad-a376-4e752e5fe5cb||174.120.8.226||2013-09-18 13:22:30||2013-09-18  
13:27:03||/hadoop-mahout-roadmap/||/hadoop-mahout-roadmap/||"-||4
```

参考

Hive综合实验(受访
与访客分析)-实验
手册

任务2 受访与访客分析

任务需求

利用数据仓库Hive完成Web日志的分析。

4. 数据集4的格式如下，各列的含义分别是：序号，年、月、日、小时。

1, 2013, 09, 17, 22

2, 2013, 09, 17, 23

3, 2013, 09, 18, 00

4, 2013, 09, 18, 01

5, 2013, 09, 18, 02

6, 2013, 09, 18, 03

参考

Hive综合实验(受访
与访客分析)-实验
手册

任务2 受访与访客分析

任务需求

5. 完成以下数据处理业务。

- 1). 创建数据仓库表
- 2). 数据导入表
- 3). 创建明细宽表
- 4). 计算各页面PV
- 5). 统计20130918这个分区里面的受访页面的top10
- 6). 统计每日最热门页面的top10
- 7). 按照时间维度比如小时来统计独立访客及其产生的 pv
- 8). 统计每小时独立访客总数
- 9). 统计每天的新访客数量。
- 10). 回头/单次访客统计
- 11). 统计每日所有回头访客及其访问次数
- 12). 统计人均访问的频次
- 13). 统计人均页面浏览量

参考

Hive综合实验(受访
与访客分析)-实验
手册

任务2 受访与访客分析

任务分析

首先创建ODS层表，ODS（Operational Data Store）是数据仓库体系结构中的一个可选部分，也被称为贴源层。ODS具备数据仓库的部分特征和OLTP系统的部分特征，它是“面向主题的、集成的、当前或接近当前的、不断变化的”数据。把数据再导入到ODS层表，再创建ODS层宽表用于统计分析。中间阶段可以创建中间表最后汇总到DW中。

任务2 受访与访客分析

任务步骤

- 1、启动Hadoop伪分布运行环境
- 2、启动Hive
- 3、创建数据仓库及表
- 4、导入数据
- 5、数据分析

任务2 受访与访客分析

任务结果

根据需求执行SQL语句，展示每个SQL的数据结果。

谢谢观看

THANKS FOR WATCHING