网站访问日志清洗

截至时间为12月19日 23:59:59

数据产生价值, 价值驱动业务. 在大数据开发项目中, 往往是需要从真实业务数据中提取出提取出有价值的信息, 称为业务指标. 从而结合业务场景, 通过业务指标、行业趋势分析当前业务在整个业务周期中的表现, 进而指导决策者对业务进行调整. 例如, 从搜狗搜索日志中分析出搜索次数排名前10的网站, 进而可以确定当前网络环境的热门话题, 这样, 企业可以投入更多资源在这些话题上, 以更少的投入换取更多的利益.

本次实践的数据来源于某个网站的后台访问日志, 每次访客访问该网站, 就会在后台日志文件末尾追加一条信息. 实践的目的就在于通过对该网站日志进行分析，计算该网站的一些关键指标，供运营者决策时参考.

---

1. 项目的输入输出

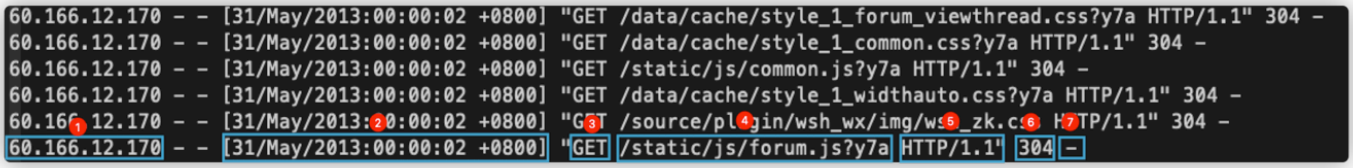
1.1 输入数据

本次实践课程总共有两份数据, 分别对应该网站2天产生的日志数据:

(1) 2013年5月30日的访问日志, 存储文件为 `2013_05_30.log` , 该文件大约记录着50W条的访问日志;

(2) 2013年5月31日的访问日志, 存储文件为 `2013_05_31.log` , 该文件大约记录着140W条的访问日志.

每个文件的内容分布如下:



其中, 每一行表示访客访问网站时产生的记录, 从左至右分别为: (1)访客的ip地址、(2)访问时间、(3)请求方法、(4)访问的网址路径、(5)网络协议、(6)状态码、(7)传输流量.

> 数据下载地址:
>
> 1. https://practice-data.oss-cn-hangzhou.aliyuncs.com/access_2013_05_30.log
> 2. https://practice-data.oss-cn-hangzhou.aliyuncs.com/access_2013_05_31.log

1.2 业务指标

本次实践课程旨在提取数据的价值, 也就是业务指标. 需要提取的业务指标如下:

1.2.1 网站浏览量

- **定义**: 网站浏览量即为PV(Page View), 是指所有用户浏览页面的总和, 一个独立用户每打开一个页面就被记录1 次
- **分析**: 网站总浏览量, 可以反应用户对于网站的兴趣, 就像收视率对于电视剧一样
- **计算**: 记录计数, 从访问日志中获取所有访问记录数量

1.2.2 新用户注册量

- **定义**: 新用户的注册规模, 该网站的登陆路径为member.php, 而用户点击注册时需要访问member.php?mod=reg的访问路径
- **分析**: 新用户注册量能反应业务中某项措施的效果. 例如, 决策者在该网站增加一项"打赏作者"的功能后, 如果增加功能后的新用户注册量比

  增加前要多, 则说明该"打赏作者"功能是正向的, 否则, 可能需要进一步的分析
- **计算**: 访问注册网址的记录计数, 从日志中获取出请求路径中包含member.php?mod=reg的所有记录数量

1.2.3 IP数

- **定义**:一天之内，访问网站的不同独立IP个数，其中同一IP可能访问了多个页面，但是独立IP数为1
- **分析**: 这是我们最熟悉的一个概念，无论同一个IP上有多少电脑、多少用户，从某种程度上来说，独立IP的多少，是衡量网站推广活动好坏最直接的数据
- **计算**: 独立IP计数，获取访问日志中不同IP的个数

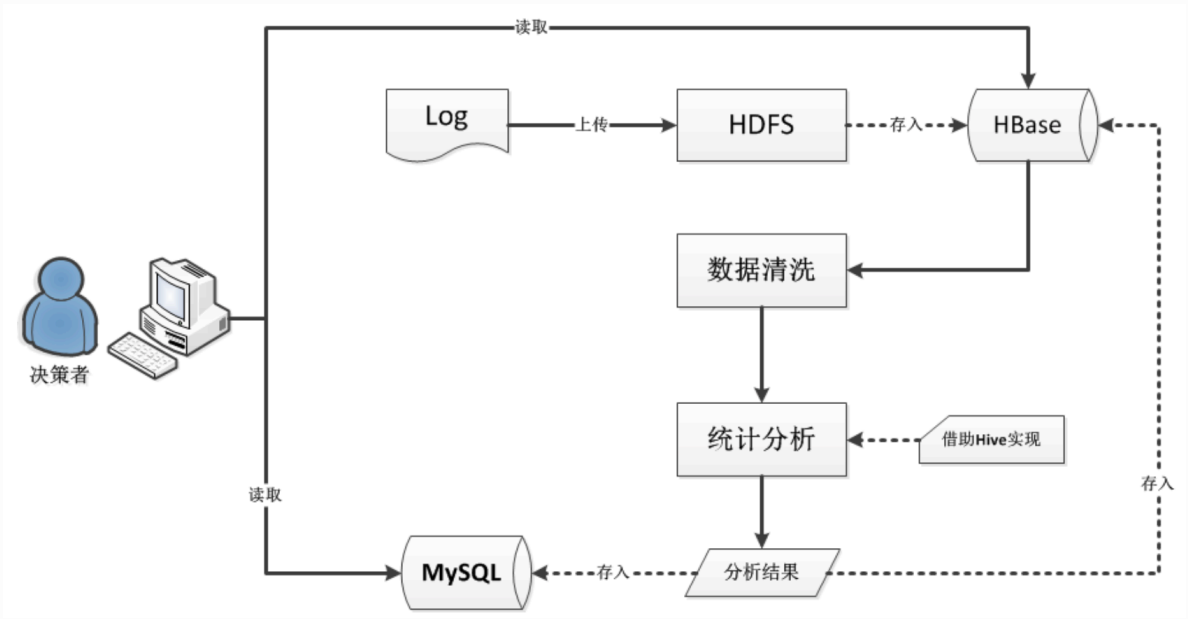
1.2.4 跳出率

- **定义**: 只浏览了一个页面便离开了网站的访问次数占总的访问次数的百分比，即只浏览了一个页面的访问次数 / 全部的访问次数汇总
- **分析**: 跳出率是非常重要的访客黏性指标，它显示了访客对网站的兴趣程度，跳出率越低说明流量质量越好，访客对网站的内容越感兴趣，这些访客越可能是网站的有效用户、忠实用户
- **计算**: 统计一天内只出现一条记录的ip的记录数与访问日志所有记录数的比值

2.开发步骤

本次数据开发过程需要用到的技术

- HDFS、MapReduce
- HBase、Hive
- Sqoop、Mysql



> 1)在HDFS的`/lessons/practice`路径下创建一个目录，作为你的工作目录，目录名称为"**姓名+学号**".
>
> 2)在你的工作目录下又创建两个目录`origin_log`和`cleaned_log`，分别存放原始数据和清洗好的数据.

2.1 上传日志文件至大数据平台

- 把日志文件上传到HDFS中进行处理，一般情况下可以分为以下几种:
  - 如果是日志服务器数据较小、压力较小，可以直接使用shell命令把数据上传到HDFS中
  - 如果是日志服务器数据较大、压力较大，使用NFS在另一台服务器上上传数据
  - 如果日志服务器非常多、数据量大，使用flume进行数据处理；

> 1)需要将数据从**本地文件夹下**导入至你的工作目录下的`origin_log`目录下.

- 把HDFS日志文件保存至HBase中:
  - HBase一般用作保存原始的结构化数据，方便后续回看数据

> 1)需要在HBase中创建一张数据表，表名为`log_姓名_学号`，并使用MapReduce将数据导入至HBase中. HBase数据表设计建议如下:

	all	**all**	**all**	**all**	**all**	**all**	**all**
rowkey	ip	time	method	path	protocol	code	volumn

2.2 数据清洗

使用MapReduce对HDFS中的原始数据进行清洗，以便后续进行统计分析

> 1)使用MapReduce清洗原始数据，并将清洗好的数据分别存放在你的工作目录下的`cleaned_log`文件夹下的`2013-05-30`和`2013-05-31`文件夹中.

2.3 统计分析

使用数据仓库Hive对清洗后的数据进行统计分析

> 1)创建Hive工作数据库 `log_姓名_学号` . 在工作数据库中完成以下任务:
>
> 2)由于数据按照日期分为多个文件, 可以考虑创建一个外部分区表存储清洗后的数据, 分区字段为当前日期, 数据表存储路径设置为上一小节的cleaned_log文件夹. 表名自定, 结构.
>
> 3)对于每个指标, 你需要创建一个数据表, 表名自定, 结构自定.
>
> 4)计算每个指标, 将指标计算结果存储在对应数据表中.
>
> 5)创建一个大的汇总表, 将所有指标放在这张汇总表中(关联查询).

2.4 分析结果存入RDBMS

使用Sqoop把Hive产生的统计结果导出到Mysql中, 最终的指标数据量一般较少, 而且前端界面需要使用指标结果, 因此需要将指标结果放入查询速度较快的RDBMS(例如Mysql)中.

> 1)在MySQL中创建一个工作数据库, 名为 `log_姓名_学号`
>
> 2)在工作数据库中创建一张指标表, 表名自定, 结构自定.
>
> 3)使用Sqoop将Hive的汇总表数据导出至指标表中.