大数据处理综合实验课程（2025）

课程设计-日志分析

南京大学　　计算机学院

1 课程设计目标

本课程设计的目标是通过Hadoop大数据编程计算平台，实现对海量日志的数据分析与数据挖掘。通过本课程设计，可以系统性学习并实践对海量日志数据的解析与清洗，浏览量数据分析，用户行为与网页请求失败关联性挖掘，用户画像与用户分群。通过任务分工与协同合作，提升项目开发、协作和问题解决能力，加深对MapReduce编程和推荐算法的理解。

2 学习技能

本次课程设计可以熟悉和掌握以下技能：

1、从海量日志中解析关键信息，清洗数据。

2、利用MapReduce,对海量日志进行网站浏览量数据分析。

3、利用MapReduce，挖掘用户行为与网页请求失败关联性。

4、利用MapReduce，基于K-means算法用户画像与用户分群。

3 题目描述

海量Web日志记录蕴含了丰富的用户访问页面的信息。本课程设计的任务是基于MapReduce 技术，实现日志数据分析与数据挖掘。如未特殊说明，要求所有任务均使用 MapReduce 程序完成，Reducer的数量要求为2~4个。本课程设计主要包括以下四项任务(详细定义见后续内容)：

1、解析日志信息，清洗数据。

2、基于日志信息进行网站浏览量数据分析。

3、挖掘用户行为与网页请求失败关联性。

4、提取用户特征，构建用户画像，实现用户分群。

3 题目描述

3.1 日志文件结构定义

本题给出的日志文件的格式：

```
222.68.172.190 - - [18/Sep/2013:06:49:57 +0000] "GET /images/my.jpg HTTP/1.1" [200] 19939
"http://www.angularjs.cn/A00n" "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/29.0.1547.66 Safari/537.36"
```

数据地址：/user/root/FinalExp/FinalExp2/dataset/access_log.txt

用于本地调试的小规模数据集：/user/root/FinalExp/FinalExp2/sample/access_log.txt

字段名	解释
remote_addr	记录客户端的ip地址，222.68.172.190
remote_user	记录客户端用户名称，- -
time_local	记录访问时间与时区，[18/Sep/2013:06:49:57 +0000]
request	记录请求的url与http协议，“GET /images/my.jpg HTTP/1.1”
status	记录请求状态,成功是200，200
body_bytes_sent	记录发送给客户端文件主体内容大小，19939
http_referer	用来记录从那个页面链接访问过来的，“http://www.angularjs.cn/A00n”
http_user_agent	记录客户浏览器的相关信息，“Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/29.0.1547.66 Safari/537.36”

3 题目描述

3.2 本题任务

任务1：对日志文件结构进行解析

解析结果可参考如下格式：

```
remote_addr:222.68.172.190
remote_user:- -
time_local:18/Sep/2013:06:49:57
request:/images/my.jpg
status:200
body_bytes_sent:19939
http_referer:"http://www.angularjs.cn/A00n"
http_user_agent:"Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/29.0.1547.66 Safari/537.36"
```

3 题目描述

任务2：网站访问分析

（1）每小时网站浏览量

用户每小时访问量统计。以每小时时间为键，每小时时间段内的访问次数进行累加即可得到结果。

请对访问量从高到低进行排序，要求程序具有大规模可扩展性。

输出格式：每一小时的时间信息[\TAB]对应该一小时内网站的访问次数

形如：

```
2013091808  2040
2013091807  1003
2013091806  111
```

（2）访问网站的用户端类型统计

根据客户浏览器信息(http_user_agent字段），统计不同用户端类型对网站的访问情况。

请对用户数量从高到低进行排序，要求程序具有大规模可扩展性。

输出格式：用户端浏览器类型[\TAB]该类型的用户数量

3 题目描述

任务3：用户行为与请求失败关联分析

根据用户行为(访问时间time_local，传输数据量body_bytes_sent，用户类型http_user_agent)，挖掘其与请求状态（status）的关系。

（1）分析传输数据量与状态码的关系

根据body_bytes_sent字段和status字段，分析请求失败与大文件传输是否相关。

输出格式：状态码[\TAB]平均传输数据量

（2）分析小时级错误率波动

根据time_local字段和status字段，绘制时间热力图，分析高请求失败时间段。

输出格式：时间段[\TAB]错误率 （按错误率从高到低排列）

（3）识别高风险用户类型

根据任务2(2)中用户人数最高的十种浏览器类型，输出不同用户浏览器类型的请求失败的错误率与平均数据传输量。分析不同用户浏览器类型导致请求失败的情况。

输出格式：用户浏览器类型->[状态码1：错误率，平均数据传输量；状态码2：错误率，平均数据传输量；状态码3：错误率，平均数据传输量] （按错误率从高到低排列）

3 题目描述

任务4：用户画像与用户分群

分析用户行为特征（如访问频率、设备类型等），并根据特征对用户分群。

（1） 获取用户行为特征

从网页日志中提取用户行为特征，并转换为数值型向量。

如访问频率（高频次或低频次）、访问时间段（白天或晚上）、设备类型（PC端用户或移动端用户）等行为特征。

（2）根据特征对用户分群

根据用户行为向量特征，使用K-Means聚类算法进行用户分群。

根据聚类结果分析每个群体的典型行为（如“移动端夜间高频用户”）。

课程设计要求

1. 提交材料

- 程序源代码，包含完整目录结构的src包，并提供编译方法说明；
- 可执行jar包以及jar包的执行方式说明；
- 程序设计报告(pdf格式)，报告内容包括每个子任务涉及的程序的主要流程、程序运行结果截图、在实现过程中进行的优化工作、优化取得的效果说明(如有)。

以上材料打包为一个zip压缩包。

- 课程设计报告(pdf格式)，报告内容包括:
  - 小组信息(姓名、学号、联系方式)
  - 小组分工情况:明确各成员在课程设计中分工的内容，要求以 git 形式分工合作(小组组长使用 git.nju.edu.cn 创建一个项目仓库，其余小组成员以开发者身份加入该项目，注意要在仓库中备注小组信息)。
  - 详细设计说明，包括详细程序设计、程序框架、功能模块、参数的选定、主要类的设计说明，包括主要类、函数的输入输出参数、尤其是map和reduce函数的输入输出键值对详细数据格式和含义，主要功能和算法代码中加清晰的注释说明，如果有优化点或创新点，请明确说明。
  - 总结:程序的特点总结，功能、性能、扩展性等方面存在的不足和可能的改进之处。

课程设计要求

2. 完成周期

- 课程设计完成周期为一个月，验收截止日期为:2025.06.04（周三）实验课结束
- 后续实验课均要求到教室
- 提交方式:上传课程网站

3. 提醒

- 设计程序时不一定要完全按照文档中给出的方法，可以有自己的解决方案，鼓励探索新方案；
- 提交材料前请务必检查好是否有遗漏、文件格式是否符合要求；
- 切勿抄袭代码；
- 程序执行说明务必详细。