Map-Reduce 课程设计 之

日志统计分析

1 课程设计目标

本课程设计通过使用 MapReduce 来实现日志分析 ,日志分析在互联网企业应用很广，通过本课程设计的学习，可以进一步了解 MapReduce 技术在工业界的应用。

2 学习技能

本次课程设计可以熟悉和掌握以下 MapReduce 编程技能：

1、 海量日志数据的统计分析

2、 基于 MapReduce 的预测模型设计，通过对历史日志数据的分析建立预测模型

3 题目描述

电商公司越来越重视接口访问日志的利用 ,从日志文件里边可以获取到接口的访问性能、访问频率、访问来源，统计有以下的意义：

1、能够快速获取接口访问性能是否下降，或者接口访问频率异常。

2、结合公司的访问量，可以预估举行促销活动时，需要增加机器的数量。

3、接口修改后，是否出现波动等。

3.1 日志文件结构定义

本题给出的日志文件的格式：

```
172.22.49.26 [16/Sep/2015:00:22:23 +0800] "GET /tour/category/query HTTP/1.1"
GET 200 156 2
```

具体意义如下表所示：

172.22.49.26	调用方的 IP
[16/Sep/2015:00:22:23 +0800]	调用的时间
GET /tour/category/query HTTP/1.1	HTTP 请求，其中/tour/category/query 是请求的 URL，URL 不同，则视为不同的接口
GET	HTTP METHOD
200	HTTP 状态码，其他经常见到的还有 404，500
156	RESPONSE 返回的字节长度
2	本次请求响应的时间，单位为毫秒

3.2 本题任务

按照以上给定的日志文件格式，按照以下的要求进行分析统计

1、统计日志中各个状态码（200，404，500）出现总的频次，并且按照小时时间窗，输出各个时间段各状态码的统计情况。统计文件命名为 1.txt 输出格式为：

```
200:100
404:2
500:1
12:00-1:00 200:5 404:1500:0
1:00-2:00 200:5 404:0 500:0
…
```

单词与数字之间英文(:)分割。时间之间英文(-)分割，其他是空格或者空行。

2、统计每个 IP 访问总的频次，并且按照小时时间窗，输出各个时间段各个 IP 访问的情况。每个 IP 的统计信息是一个文件，并且以 IP 为文件名（后缀为 txt，如：172.22.49.26.txt），每个文件的输出格式同任务 1。

3、统计每个接口(请求的 URL)访问总的频次，并且以接口为文件，按照秒为单位的时间窗，输出各个时间段各接口的访问情况。每个接口的统计信息是一个文件，如接口 /tour/category/query 的统计文件命名为：tour-category-query.txt，每个文件的输出格式同任务 1。

4、统计每个接口的平均响应时间，并且以接口为分组，按照小时时间窗，输出各个时间段各个接口平均的响应时间。每个接口的统计信息是一个文件，如接口/tour/category/query 的统计文件命名为：tour-category-query.txt，每个文件的输出格式同任务 1。

5、接口访问频次预测，给 2015-09-08.log 到 2015-09-21.log 共 14 天的日志文件，作为训练数据，设计预测算法来预测下一天（2015-09-22）每个小时窗内每个接口（请求的 URL）的访问总频次。输出格式同任务 1。该结果会与当天实际的统计值(2015-09-22.log)做 RMSE 验证。评判标准如下：

$$RMSE = \frac{1}{M}\sum_{i=1}^{M} \sqrt{\frac{\sum_{j=1}^{N}(C1j - C2j)^2}{N}}$$

其中M为时间窗个数。N为第i 个时间窗中访问 URL 个数。$C1j$和$C2j$分别是预测值和真实值第j 个 URL 的访问频次。同学们需自己实现 RMSE 公式并进行计算输出 RMSE 值，评分时会根据 RMSE 值则判定预测模型的正确度。

3.3 输入文件

输入日志文件均在压缩包 JN1_LOG.rar 内。

任务 1-4 请使用 2015-09-08.log 作为输入文件。

任务 5 请使用 2015-09-08.log 到 2015-09-21.log 的日志作为预测的输入文件，并且预测下一天每个小时窗内每个接口（请求的 URL）的访问总频次并根据预测结果和实际值计算 RMSE。

4 提交材料

请各位同学提交如下材料。

1、程序源代码，要求提供包含完整目录结构的 src 代码包，并且提供编译方法说明。

2、程序可执行 jar 包，jar 包将会以以下命令运行、评测。程序需要 5 个输入参数，对于任务 1-4 而言，第一个为输入文件路径，第二个、三个、四个、五个分别为任务 1、任务 2、任务 3、任务 4 输出目录 (例 :hadoop jar program.jar inputPath outputath1 outputPath2 outputPath3 outputPath4)。对于任务 5 而言，第一个参数为包含所有日志文件的文件夹路径，第二个参数为输出文件路径。本题目的运行环境在 hadoop-2.7、jdk-1.7 或以上环境下，必须采用 MapReduce 编程模型。

3、程序设计报告。报告内容包括程序设计的主要流程、程序采用的主要算法、进行的优化工作、优化取得的效果、程序的性能分析以及程序运行截图等。

5 友情提醒

1、请务必遵守前文所述的各种运行要求和文件格式。

2. 请附带源代码的编译说明。

3. 可先使用小数据集来调试程序的正确性，再用大数据集进行性能调优。评测时会综合考虑选手程序的正确性和性能。