数据集成综合作业

数据集成综合作业

一、作业目的

经过数据集成部分的学习，我们已经掌握了基本的数据集成的概念、方法。本次作业考察大家如何对数据进行处理，并按照指定的方式进行输出。

二、数据与工具

本次作业采用的是大家访问ftp服务器时，服务器所记录的日志记录。

1. 日志记录记录了每个同学对服务器访问的内容、时间等等信息。
2. 为保护隐私，我们已经对整体日志内容进行了ETL处理，每人只能看到自己的日志信息。
3. 每人在自己的目录下，下载名为snapper.log的文件，作为输入信息。

```
{你的ftp根目录}/snapper.log          # 注意：每个人的文件内容是不一样的
```

ftp日志具有规整的记录格式，各项含义具体说明如下：

```
Tue Nov 16 15:41:44 2021       # 记录的时间
1                              # 传输文件花费的时间（秒）
113.57.80.253                  # 客户端ip地址
437423                         # 传输的文件大小（byte）
/zhuguobin/Maze.zip            # 传输的文件
b                              # 传输类型，b表示二进制传输，a表示
ascii码传输
_                              # 特殊动作标记，可忽略
i                              # 传输方向，o表示从服务器下载，i表示
向服务器上传
g                              # 存取模式，g表示虚拟用户，a表示匿名
用户
profzhu                        # 用户名
ftp                            # 服务名
0                              # 授权方式，可忽略
*                              # 表示授权用户已认证IP，可忽略
c                              # 完成状态，c表示complete，  i表示
incomplete
```

本次作业采用的是开源的日志提取工具：goaccess。该工具有各种操作系统下的版本，请大家自行下载。

三、作业内容

本次作业的要求是：

1. 对snapper.log日志数据进行数据提取，转换为goaccess所能处理的内容，最后通过goaccess工具进行可视化输出。
2. 事实上，goaccess工具并不是对ftp日志最佳的日志提取与处理工具，它主要是针对http server日志进行分析使用的。但是，由于该工具小巧、高效（每秒可以处理数万条记录）、可视化程度高，我们可以借助它对我们的访问日志进行分析。
3. goaccess由于是针对http server日志的，因此，其分析的内容总是与http协议下的内容对应的。但是，我们使用的是ftp协议，因此，我们在作业时，需要将上面的ftp日志条目中的各项找到goaccess所能对应的表达。当然，有些项goaccess是无法表达的，如：最后一项「c」，表示的是访问成功，但是在http中，表达成功的代码是2xx（即：2打头的3位数），表达失败的是4xx，或者5xx。因此，各位同学在作业前，需要仔细阅读goaccess的默认配置文件，将其修改为可以处理我们的ftp日志的格式。
4. goaccess可以将日志文件转换为各种表达格式（如：csv、html、json等等）。它内部也使用了自己的数据库管理。我们需要仔细理解它的命令行参数，以实现自己的转换目

标。由于我们需要利用其强大的可视化能力，当然我们需要将其转化为html格式。
5. 你可以自建一个web server，以访问所生成的html，这不是强制要求的。也可以直接在浏览器中打开该html文件以可视化。

四、提交成果

提交成果要求：

1. 成果提交到你的 ftp根目录下，共5个文件。命名方式为：
    - index.html: 所转换的html文件
    - overall.png：总体分析图，即dashboard中的第一项（OVERALL ANALYZED REQUESTS）
    - visit_per_day.png：每日访问趋势图（UNIQUE VISITORS PER DAY）
    - request.png：请求文件统计图（REQUESTED FILES (URLS)）
    - time_dist.png：24小时的时间分布图（TIME DISTRIBUTION）
2. 上面的图片是可视化页面的截图，格式可以是png，也可以是其它的，如：jpg等
3. 提交的成果要按照上面的命名方式，不要自行命名。也不要对页面进行完整的长截图，而是每个内容对应一个截图提交。

五、打分规则

打分规则如下：

1. 是否完整提交上述的5个文件。html文件占60%，其余截图各占10%；
2. html是否可以在浏览器中正常打开，并正常可视化表达出来；
3. html所展示的内容，与你提交的png截图是否一致；