数据开发工程师笔试题

如无不便，请您使用 github.com 提交答案，回复 git 代码库地址即可。否则，请使用 zip 压缩包以邮件附件形式提交代码。

（一）现有 Nginx 服务器日志文件大约1000万行，其中一行如下：

```
47.29.201.179 - - [28/Feb/2019:13:17:10 +0000] "GET /?p=1 HTTP/2.0" 200 5316 "https://domain1.com/?p=1" "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/72.0.3626.119 Safari/537.36" "2.75"
```

请编写 Python 代码，统计以下数据指标：

- 计算 HTTPS 请求有多少个是以 domain1.com 为域名
- 给定一个日期 `date`，根据 HTTP 状态码计算当日（UTC时间）所有请求中成功成功的比例

（二）现有事件记录表 `event_log`，每次玩家在游戏中开启关卡，都会产生一条记录。表前几行形如：

user_id	event_timestamp
8373613	1603189321
3232343	1603189452
1343299	1603189498
8ʃ372761	1603189611
7689821	1603189734
…	…

`event_timestamp` 使用 epoch/Unix timestamp 格式
请编写 SQL 查询，查询有多少用户在2020年9月开启关卡数大于等于 1000 小于 2000。

（三）请根据以下业务规则建立国际象棋比赛数据模型：

1. 建模范围：
    - 俱乐部（Clubs）
    - 棋手（Players）
    - 会员（Members）
    - 锦标赛（Tournaments）
    - 比赛（Matches）
2. 数据关系
    - 俱乐部可以有很多会员
    - 一个棋手可以有一个且只有一个排名
    - 俱乐部可以举办许多锦标赛
    - 锦标赛也可以由其他组织赞助，（如企业或者政府）
    - 每年都有许多锦标赛
    - 棋手可以参加零次或多次锦标赛
    - 一个棋手在任何时候只能是一个俱乐部的会员
    - 一个锦标赛可以有许多棋手
    - 一个锦标赛可以有在两个参赛棋手之间进行的零场或多场比赛
3. 其他信息
    - 俱乐部由一个独特的ID识别，并有一个名称、地址和其他细节。
    - 棋手由一个独特的ID识别，并有一个名称、地址和其他细节。
    - 锦标赛由一个独特的代码识别，并有一个名称、赞助方、开始日期和结束日期。
    - 锦标赛可以有零个、一个或多个赞助方。