后端现场代码试题

掌众科技后端技术面试组

June 5, 2019

规则

1. 本环节的目的在于提供一种证明你用代码解决问题的能力的方式。
2. 以下题目中，每道题的一个小问算一个问题。现场笔试，任选至少 2 个问题，60 分钟内完成；离线笔试，任选至少 5 个问题，24 小时内完成。如果有兴趣或者游刃有余，欢迎多选。
3. 没有达到上面这条标准的，不进入下一轮面试。
4. 允许使用任何编程语言，任何可用的工具。如果你能和我们一样用 Scala 解决问题，我们会更开心。
5. 允许使用任何搜索引擎, 但不可向他人求助。
6. 以迅速、有效地解决实际问题为目标，不必拘泥或者局限于“大数据”“Spark”“Hadoop”“分布式”等概念。
7. 请体现良好的设计风格和代码习惯。
8. 每个问题的代码运行时间应该不超过 5 分钟，交付的代码必须能够直接运行得出结果。

题目

A1. 日志分析

桌面上有一个名为 `3.csv` 的文件，约有 150 万行, 其格式为:

```
id,event_type,ad_type,ip1,ip2,ext
```

1. 请按照 `ad_type` 分类统计每种 `ad_type` 的数目，并按数量倒叙排序。结果示例:

   ```
   571115 NATIVE
   482729 BANNER
   314666 INTERSTITIAL
   198562 SPLASH
   ```

   请注意，本题以本机当场算出结果为准，仅有代码而无法算出结果的解答，将被判为零分。

2. 这个文件中记录的是直接从服务器上抓下来的广告曝光和点击的日志。在广告业务场景中，广告的一次曝光可能对应着一次点击，也可能没有点击。文件中 `event_type = 7` 的行是曝光数据，`event_type = 9` 的行是点击数据，它们通过相同的`id` 关联起来(即`id` 相同的行表示一组曝光-点击)。现在我们发现，有一些暂时还未知的原因会导致一组曝光的 IP(`ip2`) 和点击的 IP(`ip2`) 不一致。请算出文件中所有这种 IP 不一致的点击，所占点击总数的比例。请注意，本题以本机当场算出结果为准，仅有代码而无法算出结果的解答，将被判为零分。

3. 如果我们服务器上有 1 万个这样的文件，仍然需要你来解决上述两个小问题，你大概会怎么做？能不能提供一种不依赖于 Hadoop/Spark 的解决方案?

A2. 幂次权重抽奖

考虑一个线上抽奖活动:

1. 一共有 n 位用户 $U_i(1 \le i \le n)$ 参与抽奖,
2. 每轮抽奖都有且仅有一名中奖者,
3. 每位用户的权重系数为 $w_i(1 \le i \le n)$, w_i 均为正整数, $w_i \le 30$,
4. 权重为 w_i 的用户，中奖的概率是权重为 w_j 的用户的 $2^{w_i - w_j}$ 倍$(1 \le j \le i \le n)$. 比如，权重为 5 的用户，中奖概率是权重为 4 的用户的 2 倍，是权重为 3 的用户的 4 倍，以此类推。

问题:

1. 请针对这样的抽奖活动，设计并实现一个算法，并对下面的用户权重运行 1000000 次抽奖，并输出每位用户的中奖数统计来检验你的算法:

id	权重
1	1
2	2
3	2
4	3
5	5
6	7
7	7
8	8

2. 如果 n 非常大，达到万级甚至百万级，并且你要处理的是一个访问量很大，并发数很高的抽奖服务，你会如何调整和优化你的算法?

B1. 拉马努金数

20 世纪初，英国数学家哈代某次乘车，注意到车牌号为 1729，便向拉马努金抱怨说，这真是一个无聊乏味的数字。拉马努金却凭借超乎寻常的敏锐直觉，指出:“这是一个非常有趣的数字。它是能用两种不同方式表示为两个正立方数之和的最小的数。”

我们把这种至少能用两种不同的方式，表示为两个正立方数之和的数叫做“拉马努金数”。

如果我们从小到大数，那么第一个拉马努金数 $1729 = 1^3 + 12^3 = 9^3 + 10^3$，第二个拉马努金数 $4104 = 2^3 + 16^3 = 9^3 + 15^3$.

请你发挥现代计算机的威力，

1. 找出最小的前 50 个拉马努金数。
2. 如果可能，请找出前 100 个。

请千万注意不要有遗漏。为了表示我们对天才拉马努金的敬意，遗漏任何一个，本题计零分。

B2. 计算自幂级数的和

定义自幂级数和为

$$f(n) = \sum_{k=1}^{n} k^k$$

前 10 项的自幂级数和为

$$f(10) = \sum_{k=1}^{10} k^k = 1^1 + 2^2 + 3^3 + \ldots + 10^{10} = 10405071317$$

记 $f(n)$ 的最后 10 位数字为 $g(n)$(0 开头的略去), 即

$$g(n) \equiv f(n) (mod\ 10^{10})$$

那么按照上面的结果，我们知道

$$g(10) = 405071317$$

问题:

1. 请算出 $g(1000)$.
2. 如果可能，请试算出 $g(1000000)$.
3. 如果你已经解决了上面两个小问题，按照你的已经实现的算法，试估算，计算 $g(10^{10})$ 大约需要多长时间。

B3. 斐波那契数列

在数学上，著名的斐波那契数列以递归的方法来定义：

$$F_0 = 0$$
$$F_1 = 1$$
$$F_n = F_{n-1} + F_{n-2} \quad (n \geq 2)$$

1. 试算出 F_{100} 的准确数值。
2. 请问斐波那契数列中，第一个有 1000 位数字的是第几项?
3. 试算出第 10 亿项除以 1000000007 的余数，即 $F_{10^9} \ mod \ (10^9 + 7)$.
4. 试算出第 10^{1000} 项除以 1000000007 的余数，即 $F_{10^{1000}} \ mod \ (10^9 + 7)$.

B4. 乘幂的数字和

Googol(10^{100}) 是一个超大的数，写出来是 1 后面跟着 100 个 0. 100^{100} 则更是大得丧心病狂：1 后面跟着 200 个 0. 然而，尽管这两个数如此巨大，各位数字和却都只有 1. 而 $7^{10} = 282475249$, 其各位数字和为 43.

那么，对于自然数 p, q 满足 $0 < p < 100, 0 < q < 100$，所有能表示为 p^q 的自然数中，最大的各位数字和是多少?

B5. 字典序排列

排列指的是将一组物体进行有顺序的放置。例如，3124 是数字 1、2、3、4 的一个排列。如果把所有排列按照数字大小或字母先后进行排序，我们称之为字典序排列。0、1、2 的字典序排列是：

012　021　102　120　201　210

1. 数字 0、1、2、3、4、5、6、7、8、9 的字典序排列中第一百万位的排列是什么?
2. 现有以下

- Apple 苹果,
- Betelnut 槟榔,
- Carambola 杨桃,
- Coconut 椰子,
- Durian 榴莲,
- Ginkgo 银杏,
- Lichee 荔枝,
- Mango 芒果,
- Nucleus 核仁,
- Olive 橄榄,
- Pear 梨,
- Watermelon 西瓜

共 12 种水果，按其英文作字典序排序。

假设我们已经成功地求出了所有的排列，现在需要构建一个 Web 服务，接受全球各地许多用户输入的排列序号（比如说 1000000 位），输出对应的排列。

请你设计一个后端的解决方案，越详细，越能体现你的水平，越好。