Facebook on-site:

1. Data Model

Design Uber (focus on data model). Select rides from and to airport only.

2. Photo upload

Actions/steps:

login (has session\_id). 1point3acres

click on post

upload photos

tag friends

upload photo again

type message

click on post

SQL question: find the time taken between each steps/actions above. For duplicate actions (upload photo again, use the first upload photo time)

Code question: same as above sql question, but input data like below [session\_id, action/step\_id, start\_ts].本文原创自1point3acres论坛

data=[['A', 1, 100]

, ['A', 2, 110]

, ['A', 3, 120]

, ['A', 3, 130]

, ['A', 4, 150]

, ['A', 5, 170]

, ['B', 1, 100]

, ['B', 2, 110]

, ['B', 3, 120]

, ['B', 3, 130]

, ['B', 4, 150]

, ['B', 5, 170]]

find the time taken between each step

3. DAU

SQL question. Given table: user\_id, last\_login\_time, prev\_login\_before\_last\_login

Find: yesterday's DAU, last 7 days, last 28 days

Coding:

users have more than one devices, last 7 days login (1 = login with that device; 0 means didn't login with that device)

details = {'iPhone':[0, 1,0,1,0,1,0]

,'Android':[1, 0,0,0,0,0,1]

,'Web':[0, 0,1,0,0,0,0]}

rollups= {'overall':['iphone', 'Android', 'Web']

,'Mobile':['iPhone', 'Android']}

-google 1point3acres

Question: find last 7 days for overall , mobile, etc

Return:

{ “overall” : [1,1,1,0,1,1,1]

“Mobile”: [1, 0,0,0,0,0,1]

…..

…

}

E.g: login session\_id A, 做了以下的actions/steps：

starttime：10:00 ， upload photo

Start time: 10:04, tag friends on photos

Start time: 10:05, upload photo again. From 1point 3acres bbs

Start time: 10:07, add messages

Start time: 10: 09, post photos (完成了）

session\_ID B （另外一个人），做类似的actions。平均每个action 多长？

为了运算的便利， start time 都是以整数来代替。

Start time 100,

Next start time 110, etc

rollup()

{'Mobile': [1, 1, 0, 1, 0, 1, 1], 'overall': [1, 1, 1, 1, 0, 1, 1]}

第三个DAU部分考了具体的new user，retained user，churn user

补充内容 (2018-10-31 07:43):

设计ride-sharing 那个要求写了这些query：计算所有从机场出发和去机场的driver；计算所有只去机场和从机场出发的passenger（乘客可能还有别的trip到别的地方这样就不算在内）讨论了一些table设计的优劣性

Facebook behavioral questions:

介绍一下你自己.本文原创自1point3acres论坛

最Challenge的project

如何处理同事Conflict

有没有同事或老板给你Critical comments, 之后如何处理

别人东西Delay怎么办. 围观我们@1point 3 acres

Project一开始没方向该怎么办

有External Obstacle怎么处理

最有趣的Project

想去什么组

为什么想加入FB

. more info on 1point3a

FB on-site:

如何evaluate一个photo upload feature是不是好，用什么metrics，收集什么数据，写了简单的query和coding，我用的Python，大概就是dictionary update

面试围绕解决实际工作中的问题。

1． 国人。 让你说出一些指标。 我提了日/周/月积极用户。 让我开始画个时间序列。

给我一个用户记录的表。 日期， 用户名， 活动。 活动是登录。 设计一个发日/周/月积极用户的表。 我开始用三个sql 来做。 但即使 扫30 天的数据也很多 。

我使用 一个sql 把3个数字算出来。 但也需要30 天的数据。 最后他提示设计一个每天用户最后哪天登陆。最后突然来了个Leetcode 78。

2． 白。 画不同设备（ 手机，电那） 我在这花了一些时间。 要建一个表。

计算人日均帖数。

3． 在大食堂吃饭。 国人很帮忙。 白人爱理不理。

4． 计算电影前x排名。 电影看多少就算。 如百分就十。 Leetcode 56。把看的分钟数加起来。

中间创建多对多关系的表。

我用了Windows function算排名。

感受

不要套近乎。 直奔主题。 时间可能不够。 不用解释很多。 你先画。 不多都有提示。

1，给你一个senario,你来想那些data 可以用来分析这个结果（比如active user, 需要看open apps time, clicks time）. 1point3acres

2，然后你再来想，最后出来的report 图表需要有哪些factor，图表怎么画， 才能最直观反映出数据增减。

3，告诉你raw data大概会是什么样的（网站/APP的 log）， 怎样设计ETL pipeline，把data倒入你最终想要的table 给report用，基本步骤：

raw data table ->intermedia table -> data warehouse fact table

4，optimize你的Process， 如果数据量非常大，如何有效地抓取数据，减少数据量（比如raw data table一天有billion data,

如何在data warehouse缩减到Million或者更少）

5， 那么问题来了，SQL怎么写join什么来得到最终table（白板）

SQL就这么多，JOIN, CASE, FUNCTIONS， （这次连subquery都不让用了，禽兽）

6，那么中间遇到的一些问题，可否用python 实现，大概考了：（白板）

[5,20],[15-35],[40,80], 缩减到overlapping [5,35],[40,80],然后返回 35-5+80-40=70

[a,b,c], enumerate: [],[a],,[c],[ab],[ac],[bc]........

根据input, dynamic 地生成SQL query.

This solves it for a static stream but we can modify the strategy to compute the running new average time spent on

each step for every new stream of data coming into the system.

import datetime

Z0

Python Code:

from collections import defaultdict

def avg\_time(streams):

result\_dict = {}

time\_spent = {}

users\_on\_steps = {1:0, 2:0, 3:0, 4:0}

for i in range(len(streams)):

stream = streams[i]

session\_id, user\_id, step\_id, current\_step\_start\_time = stream

if ((session\_id, user\_id, step\_id) not in result\_dict):

result\_dict[(session\_id, user\_id, step\_id)] = current\_step\_start\_time

if step\_id &gt; 1 and (session\_id, user\_id, step\_id-1) in result\_dict:

users\_on\_steps[step\_id-1] +=1

previous\_step\_start\_time = result\_dict[(session\_id, user\_id, step\_id-1)]

if (step\_id-1) not in time\_spent:

time\_spent[step\_id-1] = (current\_step\_start\_time - previous\_step\_start\_time).total\_seconds()

else:

time\_spent[step\_id-1] += (current\_step\_start\_time - previous\_step\_start\_time).total\_seconds()

print(result\_dict)

print("Users in each step", users\_on\_steps)

print("Time spent in each step", time\_spent)

for step in time\_spent:

avg\_time = time\_spent[step]/users\_on\_steps[step]

print("Avg. time spent on step {} is {} seconds".format(step, avg\_time))

avg\_time(streams)

O/p:

Users in each step {1: 3, 2: 3, 3: 3, 4: 2}

Time spent in each step {1: 1920.0, 2: 1920.0, 3: 840.0, 4: 300.0}

Avg. time spent on step 1 is 640.0 seconds

Avg. time spent on step 2 is 640.0 seconds

Avg. time spent on step 3 is 280.0 seconds

Avg. time spent on step 4 is 150.0 seconds

Modeling log files into useful dashboards.

How to rate the popularity of a posted video online.

how would you model Linkedin? what would be the tables and connections between them.

given a list of tuples of movie watched times, find how many unique minutes of the movie did the viewer watch e.g.

[(0,15),(10,25)]. The viewer watched 25 minutes of the movie.

t = [(0, 10), (15, 25), (12, 20), (30, 48)]

total = 0

for i in t:

total = total + i[1]-i[0]

print("total time watched:" , total)

Given a multi-step product feature, write SQL to see how well this feature is doing (loading times, step completion %). Then use Python to

constantly update average step time as new values stream in, given that there are too many to store in memory.

For SQL there are steps 1,2, and 3, and a time stamp. You can deduce time spent on each step By subtracting

timestamp 1 from 2, then averaging the differences for average time spent on each step.

For python, as new times stream in for each user (user iD and step number and timestamp) find out if we’ve

already seen this user and if so find the difference between step time stamps, add the difference to a running total,

increment the “seen” users count by 1, and divide total by number of users to be able to return the average page

loading times on demand

多准备etl 的东西。 注意他们的数据很大。 尽量东最少的书据。 思路一天为单位。

context of excercise could be a funnel analysis, how do you calculate average time bw steps

系统设计我看了某章的课程和这个课程 https://www.educative.io/collect ... Name=Design%20Gurus

FB 系统设计比较高频的几个题好像是：design POI， design instagram, design facebook messager, design news feed.

3. LC 28

4. LC 341

5. Given an input integer list and an int N. Return a list size of N which the elements in the result list is chosen from the input list randomly

FB onsite:

李特扣 yaojiu

LC erqisan

SQL，第一题判断正负问题，用CASE解决；第二题，类似于LC 要把伞，只不过有另外一个table里面是date（一月一号，一月二号等），要求找到每一天有哪些顾客没有买东西；第三题，yaoqiqi

行为问题，狂轰滥炸

设计pipeline，大意就是根据用户观看netflix电影的log，处理数据导数据库，用来支持前端的查询比如top k viewed movie。大概就是，根据log format，怎么transform数据，

然后放到数据库中什么schema；主要就是考察ETL design，和很简单的算法，最后还问了一下如何分布式处理比如用MapReduce或者Spark，总体不难

. from: 1point3acres

吐槽一下，为什么我看的面经coding题目都是easy或者combination这种题目，到我这里第一题当头一棒的DP问题，第二题又是edge case比较多的DFS，好在最后做出来了

最主要的方面就是data source -> transformation rules + source target mapping -> target data destination，中间的tranfromation主要就是做aggregation，

就是怎么把数据总结归纳起来

1. Where do you see yourself in five years?

回答1：I see myself growing into a experienced software engineer. 太泛!

回答2：我有frontend, backend的经验，我希望以后成为一个full stack. 成功安利你懂得不少。

回答3: Become a superviser and use my skills to support others. 显得有team leader的好品质。

另外，根据你申请的公司，如果你知道他们内部升级的机制，可以说两句，比如 Google 的master new grad 一进去是Level3,你可以问问身边的朋友看看一般多久久能升一级。

回答这个问题的时候，就可以带一句你对公司升级机制的了解，希望你2年之内升一级（举例子，我并不知道是不是2年可不可以升啊，不要被我误导）。

来源一亩.三分地论坛.

2. What are your weaknesses?

What are the areas where you need to improve your skills?

Are there areas where you need to develop your skills further?

这类问题叫做negative question。如果回答“I don't have any weakness" 只能说你真不懂面试官意图, 可如果回答 “My greatest weakness is I am working too hard”，这就是在搞事情了。作为一个New grad, 我的weakness就是缺少中作经验，但是答到这还没完。对于这样的negative question，当然要有转折。接下来就要说这不是个问题，我有internship experience, 做过很多projects, tech skills上和team work都有经验，不是个大问题，再来句自己是个fast learner。包括正在申请的同学，如果你想转专业，这个也是个很好的回答方式。如果不是new grad的前辈，可以看看下一题的回答。

关于这个问题的Youtube小视频：https://www.youtube.com/watch?v=4plu0nLyb\_s

3. What do you like least about your previous job/internship?

What do you dislike the most about your last job?

这个就非常尴尬了，是不是觉得这个问题很少问到。我想说的是，我onsite就被问到了这个问题，答的支支吾吾，简直要撕卷。首先，切记！切记！千万不能说之前雇主，同事的问题。即使心里有一万个抱怨，即使是在你看来微不足道的问题，只要是有一点点说 别人的不是，都不是个特别好的回答。没有面试官希望看到雇佣你以后，如果之后你再去面试其他公司，说现在公司和同事的不好。所以这题要怎么答呢。楼主看了个Youtube视频，颇受启发 https://www.youtube.com/watch?v=cK\_KUK1fDc8 以下简介概括。

换个角度，回答这个问题：What do you wish you had done differently at your previous company?

这个问题在未来可以非常容易改正. The problem can be easily changed in the future.

What you have learned from the previous experience.

粗暴的来个例子。

作为了new grad/ intern，希望自己involve到很多的task，所以对这个prioritization做的不是很好。但是知道这是不对的，我现在也慢慢做了调。之前网上找到的，供参考。

I have had problems in the past with taking on too many projects and then not delivering quality and, in some cases, not delivering on time. I was simply stretched too thin in too many areas with not enough time to deliver in a quality manner. I think that's the opportunity that is there as a new college student, you want to do everything and be involved in everything. However（敲黑板）, I've learned that I'm not at my best when I have too many conflicting priorities. I've had to cut back on some of the less important extracurricular activities to focus on delivering my academic projects with the highest quality. This prioritization of my work has carried over to my work life with my recent internship. I had several key deliverables that were due the same week, so I met with my boss to prioritize the delivery schedule. She helped me focus on both the timing and the quality of the delivery. In the end, all three projects were delivered with high quality results.

. 1point 3acres 论坛

Reference: https://collegegrad.com/tough-in ... r-greatest-weakness 这个问题和之前weakness的问题很相似，根据自己的经历，准备好问题都可以变换的回答。

准备Behavior的问题，按照第一段链接里的16个问题，再准备下所谓的negative quesitons就稳了。网上一搜behavior的问题很多很多，但很多都是大同小异。楼主面了无数个面试了，根据自己的经验下面列一下必考题和高频题。

必考题

1. Tell me about yourself

2. Why ..Company?

3. What is your most Challenging project

高频题

1. Describe a situation where you exceeded expectations/did more than required

2. Have you ever had a conflict with the team? How was it resolved? (最好举例)

3. 你想去哪个组，哪个产品，对哪方面感兴趣。当然，如果是面试特定的组，就不会问这个问题。

FB onsite:

面试过程：. 1point 3acres 论坛

1，给你一个senario,你来想那些data 可以用来分析这个结果（ 想象一下已有一个log 需要哪些必备的数据）

2，然后你再来想，最后出来的report 图表怎么画， 才能最直观反映出数据增减。（重点在时间和增减上）

3，告诉你raw data大概会是什么样的（网站/APP的 log）， 怎样设计一个intermediate table　可以减少你的数据量，raw data table一天有billion data,缩减到Million或者更少）

4，SQL怎么写join什么来得到最终table（白板）

一定注意不要向我一样见着什么就想group by 会很悲剧的

5，那么中间遇到的一些问题，用sql很困难可否用python 实现，大概考了：

　　Merge Interval .本文原创自1point3acres论坛

. 牛人云集,一亩三分地

总体来讲三轮类似的interview.

强烈建议大家试一试怎么用线性图来做，然后怎样存储你的table structure

重点来讲和System Design 很类似，建议刷类似的面经

接下来，面试官要求用图表显示data，一般都是用Line， Bar，最多用到Pie，外面流行的Data Visualization tool比如Tabulae里面fancy的图表都用不到，data visualization这一块考的很“朴实”。

后来，面试官也跟我说，facebook内部大部分用自己开发的可视化tool，足够满足平常的需求了。

-google 1point3acres

和面试官讨论清楚dimension和最终的chart之后，面试官要求进行data modeling和build ETL pipeline。这里的ETL pipeline和我平时工作中pipeline的定义差的有点大，

不知道是不是因为只是interview的关系，这里的ETL步骤简单很多，基本上一到两步就可以完成，其中SQL也没有那么多复杂的逻辑。这可能跟facebook data engineer的特点有关，

据面试官说，data engineer每天coding并不是最重要的，而是和product多打交道，多多深入了解product，这也是为什么data engineer是属于product，而不是engineer部门。

.1point3acres网

三轮面试下来，基本感觉就是FB的data engineer更加偏analysis的工作，对product也有很好的sense很重要，但是有要求有一定的coding能力。希望能够给到地里对该position感兴趣的

同学些有用的信息

碰到一个要merge interval的问题。算是考了算法吧。这个position关键不在算法，而在于product sense 和 data modeling

FB financial earning PDF:

https://s21.q4cdn.com/399680738/files/doc\_presentations/FB-Q4'16-Earnings-Slides.pdf

lc 28, 47, 65

Data Engineer 约了两轮，一轮sql，一轮code

一、 sql

sql的题没抄下来，一共6个问题，第5个问题有两小问， 用到了很多group , having, subquery. 一亩-三分-地，独家发布

能想起来的：

1. Customer table中男女比例

Customer

id

gender

2.sale table 中order size超过50个的产品数.本文原创自1point3acres论坛

Sale

sale\_id

customer\_id. visit 1point3acres for more.

product\_id

order\_size. visit 1point3acres for more.

3. 有订单的customer占所有customer中的比例

二、coding. from: 1point3acres

一共7道题，input 均是arraylist， 最后两道题没时间写完T-T

1. 最大值

2. feq 最大的数

3. median

4. 最小绝对值差. more info on 1point3acres

5. 第一个不重复值

6. 最大公约数. visit 1point3acres for more.

7. 最小公倍数

三哥已经对我很仁慈了，只考了我排序，碍于我算法实在太生疏。第三个卡住在swapping的状态转换, time out。

INPUT. 一亩-三分-地，独家发布

0 1 2 3 4

-----------------------------

IN E A B D C

O 4 0 1 3 2

. From 1point 3acres bbs

EXPECTED OUTPUT

0 1 2 3 4

------------------------------

IN A B C D E

第三个方法要求O(n), inplace.

考了SQL, ETL design, dimensional modeling design, coding..1point3acres网

SQL 具体问题记不清楚，考到了cross join。. 牛人云集,一亩三分地

coding leetcode 原题，3sum，期间改动了一下要求，出现重复怎么办.留学论坛-一亩-三分地

design 穿插一题combination

[1,2,3,4..]

[5,6,8,9...]

[8,12, 19,..]. more info on 1point3acres

答案是[4,8]

第二轮电话面试，一个等差数列，中间少一个数字，找出这个数字

follow up 1: 如果是等比数列呢？

follow up 2: 如果给一个数列，并不知道是等差还是等比，找出这个数字

这一轮表现比较突出，据说面试官给的反馈各种给力，然后就拿到了onsite

第一轮onsite(coding): 大概是先写一个SQL解决一个问题。然后给一个list of parameters to group by，自动生成一个list of SQL queries。

没什么算法，更多的是如何用最短的时间最短的代码写出来。我用的python，很多functional的语句应该加分了不少。

第二轮onsite(SQL):

然后他出了另一道题，感觉还简单一点，就是一个数组，附加一个index array表示这个数组每个数字原来应该在的位置，然后需要还原这个数组。当然in place是最好的。

比如，现有的数据是table A, B, ... 最终要显示在dashboard上的数据是从tableX来的，那么ETL pipeline就是如何从table A, B, ... 变成table X

think about daily data loading vs batch loading and scale

t = [(0, 10), (15, 25), (12, 20), (30, 48)]

t = sorted(t)

tol = t[0][1] - t[0][0]

for i in range(len(t) - 1):

if t[i+1][0] &gt; t[i][1]:

tol += t[i+1][1] - t[i+1][0]

else:

tol += t[i+1][1] - t[i][1]

print(tol)

when to solve performance by creating index and when to use the summary table

perform merge-sort with sql

query a many to many relationship while not violating the grain of a fact table(multiple-pass sql with With)

How could you handle large amount of Data during ETL?

Learn recursive functions in python

Interview Questions

What do you enjoy about your role?

Answer Question

What (soft skills ) did it take for you to get to where you are?

1 Answer

What do you think it takes to build a great Date Engineering team / Platform

1 Answer

What's the most difficult thing you've had to do?

第一轮是behavior，印度小姐姐。自己刚进去就面behavior，感觉很紧张，一些情形说得不清楚，小姐姐问了我好几处，让我clarify。 感觉这一轮是面得最不好的，之前准备都是把这些问题的答案

放在心里过一遍，和跟别人说清楚确实不太一样。然后自己resume里面有一两段经历准备得不是很充分，碰巧又被她拿出来一直问 （比如她问我一段big data project经历里面有什么和队友不合的事情，

我之前准备这个问题的时候是用另一段经历来回答，这一段就没有准备到）。还有最后一个问题，是对FB的产品有什么建议，哪些觉得可以改进的。这个我之前完全没有准备，

犹豫了一会，她说时间也到了，就这样吧

第二轮是project sense，一个中年白人大叔，一开始问一些data visualization的东西，然后让我找一个评价steps in funnel里面的metric，他期待的答案是conversion time，

我愣是没说上这个，然后他就直接告诉我了，然后基于这个metric，让我写了一些SQL和Python (Python是用来把stream改成real time的)。 后面倒是写得很顺利，

唯一不足的可能就是变量定义得不是很清楚。我用string.split 变成一个list，然后后面直接用 list[0], list[1], list[2], list[3] 他检查我code的时候，

在这种变量分别代表什么的时候有点晕，跟他解释了才清楚。

中间休息，我的HR来问问我情况，我说behavior面的不好，她一直安慰我，说后面面得好就行了，也给了我一些有用的面试经验。然后一个白人小哥带我去吃午饭，他很厉害，

是个team的manager了，和我聊了很多FB文化，职业发展的东西，我也问了一些他们正在做的项目。这个部分不算分，所以就随便聊聊了。

第三轮是 SQL 和 算法， 一个印度小姐姐。大概就是给一些背景材料，让估计Daily active user的各个方面，比如哪些是连续用户，哪些是返回用户，哪些假设已经churn 了，

然后用SQL 来计算每天的这些数据。她一步步带着我写，虽然最后的SQL还挺复杂的，但是她一步步带着我，循序渐进，所以也并不难。最后也是把这些东西用Python来写一下，

处理一些dictionary, nested list, 几个嵌套循环就出来了。感觉她也很满意

第四轮是 data modeling， 一个白人小哥。我感觉我就是刚来的时候很紧张，然后之后越来越好，最后一轮的时候已经完全放松了。这一轮让设计一个类似于Uber的 FB ride。

主要是设计后台数据库的schema，我用的是star schema，一个大的fact table连一些具体的dimentional table。感觉非常顺利，小哥人也很nice。然后基于这个设计，写了一些SQL，

也是他带着我循序渐进，写得也非常顺利，沟通也很好。最后写出来的SQL还挺复杂的，用了不少 电面考到的 having + aggregation function的 方法。自己中间稍微有几个typo，

小哥一指出我就发现了。最后还有一点时间，用来问他几个问他，以及我之前一个关于推荐系统的project（因为小哥就是推荐系统组的）。 最后walk out。

BQ：中年三姐，自我介绍、conflict with coworker、入职90天的计划

Data Modeling：中年三哥，如何evaluate 一个ride share app是不是好，用什么metrics，建什么样的data schema，写了几个query，基本只用到了join、groupby

Scenario 1: 年轻国人小姐姐，如何evaluate一个photo upload feature是不是好，用什么metrics，收集什么数据，写了简单的query和coding，我用的Python，大概就是dictionary update

Scenario 2：白人大哥，如何monitor FB DAU、WAU，给了一个table，写query找new user、retained user、resurrected user，然后时间就到了。。猜应该后面还有coding部分

总体感觉technical部分挺简单的，SQL熟练了就行了。 product sense还是要多看，对于new grad这部分也不太好准备。面完感觉整体应该过的概率比较大，主要担心就在behavior上

（因为听说FB的DE还是很重视 communication的），product sense也就conversion time没有答出来，别的也还好，后两轮感觉很不错。结果没想打今天还是受到了拒信。可能面得简单，

所以大家表现得都不错吧。正在向 HR 要feedback中。 楼主12月就毕业了，感觉现在非常慌，还没有一份offer。祝各位一切顺利，早日上岸吧

1. 有一个senario 你要考虑什么measure来回答这个问题，怎么细分可以让问题结果更精确。这块应该是考察business sense 好好研究earning report没错的。

2. 怎样构建一个图表可以显示你要解决的问题，如何考虑数据的增减比较，这个只要使用基本的bar chart/line chart/pie chart就好了。要对这些图表的使用方法很熟悉，建议好好准备因为肯定会问。

3. 给你一个原始数据表格，你需要另外构建什么表来呈现最后的dashboard，如何写SQL来得到这个表。感觉他们想要看到dashboard直接从一个physical表中读数据完全不希望看到每次用户click

的时候再去join table计算数据。

考虑到数据体量这个approach是很合理的，所以好好考虑你的最后表格是什么样的structure。

4. 数据转换的时候怎么用coding解决一些问题，考了allSubset 和 merge intervals。感觉FB这个面试不换题啊，和别人面经里面一模一样。因为第一个coding写太顺利了被问到是不是之前做过，

还能咋说，这么基础的题目可能没做过吗？

BQ：中年三姐，自我介绍、conflict with coworker、入职90天的计划

第二轮白人小哥，求DAU，但是变着花样玩，给你今日用户log和累积到昨日的dau，求merge成新的dau，楼主傻眼。白人小哥给我写好了full outer join,最后楼主憋了出来，小哥鼓励说还不错

午饭印度小姐姐，人不错，就是一直问问题，我只是想一个人静一静。。。静一静。但是一直问。。。楼主觉得挂点就在下午的第一轮

第三轮挂点轮，senior 大叔。给一个上传照片的app，怎样可视化各个维度的上传时间。之后接一个sql，求平均上传时间，秒了。还剩20分钟写code，给一个stream data，求平均时间。很简单的题目，但是很困，最后居然没写完。

跟面试官说了思路，面试官说他知道我在做什么，还说很好。但是楼主注意到他没有拍照（你的每一题都会被照下来）。

第四轮国人阿姨轮，做data modeling。接sql。没什么大问题。

BQ：中年三姐，自我介绍、conflict with coworker、入职90天的计划

Data Modeling：中年三哥，如何evaluate 一个ride share app是不是好，用什么metrics，建什么样的data schema，写了几个query，基本只用到了join、groupby

Scenario 1: 年轻国人小姐姐，如何evaluate一个photo upload feature是不是好，用什么metrics，收集什么数据，写了简单的query和coding，我用的Python，大概就是dictionary update

Scenario 2：白人大哥，如何monitor FB DAU、WAU，给了一个table，写query找new user、retained user、resurrected user，然后时间就到了。。猜应该后面还有coding部分

面试内容侧重product sense, data modeling, SQL和简单coding。Product Sense基本上就是理解你的数据的物理意义，和怎么将这些数字和公司收入联系在一起。准备的话看下公司财报里的METRICS， 如果赶上earning season，可以听听conference call反正都是免费的，看看分析师在QA环节里的提问，想想你怎么答。或者看看conf call 的 transcripts. 书的话，还是那本大家推荐的 Lean analytics，外加地里面DS的产品面经。

Data modeling 就是看Kimball 那本书, 时间短先看前两章，有个SENSE。另外用书里的方法练练怎么model 几种不同类型的公司比如 ebusiness, social media, online streaming, car sharing。

SQL coding 是重点，hacker rank + leet code + SQL cookbook 外加DS面经，总之就是知道怎么用来算metrics。

CODING 就刷题吧。。。

刚刚面完回到家。签了NDA，不能说具体面了什么。只能说地里的面经已经很全面了。Full stack interviews，clarify the requirements, gather metrics, come out dimensions, draw charts.

=> ETL from raw logs (table schema draft, SQL queries) => python/java coding. 就是地里说那样。准备好recruiter给的资料，想清楚地里已有的问题，基本不会被surprise到。

但是。。。。我仍然被surprise了。。。。第二轮做不太好，感觉面试官一直在给hints，他越给我越紧张。。。最后一轮更是震撼。。。那个面试官实在是不友好，他脑海中有个思路， 而且非常strong opinioned。我考虑到一些问题，然后给他交流说我要做一些assumption。他说好，然后我基于我的assumption给出解题思路，然后他表示很惊讶，好像我没有考虑到一些问题。。。。。而且他的language很offensive。。。有些时候不太清楚他要什么，

交流起来有点难。他问我，你到底弄没弄过data modeling。。。你知不知道什么是left join。。。。。我真的是很无语。。。

Onsite一共五轮：

第一轮：白人面试官，说是System design，其实就是ETL Design。就是从最开始获取一个需求，到最后得到结果，中间如何去进行，和面试官讨论了所用到的一些技术栈，数据库Schema的设计，

到ETL具体怎么写，怎么部署。之后面试官改动了一些需求，增加了一些功能，怎么在原来设计的基础上进行变动。总体不难，相比常规的系统设计要容易不少，估计有这方面工作经验会更得心应手一些。

第二轮：墨西哥面试官，Coding。一道大题，三小问，第一小问Easy难度，考点HashMap。第二和第三小问Medium难度，考点DFS/BFS和HashMap，其中第二和第三小问相关性比较大。之前在LC上并没有

见到过类似原题，面经中也没有出现过~

第三轮：亚裔面试官，Behavior，面试官在西雅图，是视频面试，基本上问的都是常规问题，面试之前Recruiter会发一些prep material，基本上都含盖了这轮的问题。

第四轮：一位台湾面试官，一个他的Trainer是ABC，SQL，一道大题，三小问，第一小问Easy难度，考点在使用Count()函数和Distinct关键字，第二和第三小问考点在Time window上，相关性比较大，

难度大概是Medium到Hard的难度，逻辑上不复杂，只不过实现起来更复杂一些，要多个join，还要用到sub queries。

第五轮：印度面试官，Coding，一道Easy题，warm up，LC里FB高频题之一，考点在Array。第二和第三道题相关性比较大，也都是LC原题，面经里也有，考点在Trie和recursion, 难度Medium。

总结：

从电面到Onsite，这个职位在算法上考的都不是特别偏或者特别难(也有可能是楼主运气比较好~)，

之前准备Onsite的时候，recuriter告诉我说一般LC里Medium和Hard难度的题刷个140到160道就差不多了。这个职位有一轮SQL面试，楼主把LC上四十多道SQL都刷完了，

但是面试时考到了Time Window，这个考点在LC上并没有见到，平时练习时还是要多涉猎一下~

李特扣 91

LC erqisan

SQL，第一题判断正负问题，用CASE解决；第二题，类似于LC 要把伞，只不过有另外一个table里面是date（一月一号，一月二号等），要求找到每一天有哪些顾客没有买东西；第三题，yaoqiqi

行为问题，狂轰滥炸

设计pipeline，大意就是根据用户观看netflix电影的log，处理数据导数据库，用来支持前端的查询比如topk viewed movie。大概就是，根据log format，怎么transform数据，然后放到数据库中

什么schema；主要就是考察ETL design，和很简单的算法，最后还问了一下如何分布式处理比如用MapReduce或者Spark，总体不难

吐槽一下，为什么我看的面经coding题目都是easy或者combination这种题目，到我这里第一题当头一棒的DP问题，第二题又是edge case比较多的DFS，好在最后做出来了

最主要的方面就是data source -> transformation rules + source target mapping -> target data destination，中间的tranfromation主要就是做aggregation，就是怎么把数据总结归纳起来

比如product manager要看 the number of daily/weekly/monthly active user。然后，你要考虑其中涉及的dimension包括哪些，这里就要提ER了，其实dimension就那么几个，

都在facebook的ER里提到了，比如user's country, age, gender, user's devices. 之后的两轮interview基本上也是按照这几个dimension进行分析。

接下来，面试官要求用图表显示data，一般都是用Line， Bar，最多用到Pie，外面流行的Data Visualization tool比如Tabulae里面fancy的图表都用不到，data visualization这一块考的很“朴实”。

后来，面试官也跟我说，facebook内部大部分用自己开发的可视化tool，足够满足平常的需求了。. From 1point 3acres bbs

和面试官讨论清楚dimension和最终的chart之后，面试官要求进行data modeling和build ETL pipeline。这里的ETL pipeline和我平时工作中pipeline的定义差的有点大，不知道是不是因为

只是interview的关系，这里的ETL步骤简单很多，基本上一到两步就可以完成，其中SQL也没有那么多复杂的逻辑。这可能跟facebook data engineer的特点有关，据面试官说，data engineer每天

coding并不是最重要的，而是和product多打交道，多多深入了解product，这也是为什么data engineer是属于product，而不是engineer部门。

三轮面试下来，基本感觉就是FB的data engineer更加偏analysis的工作，对product也有很好的sense很重要，但是有要求有一定的coding能力。希望能够给到地里对该position感兴趣的同学些

有用的信息

面试围绕解决世纪工作中的问题。

1． 锡金邻国人。 让你说出一些指标。 我提了日/周月积极用户。 让我开是画个时间序列。

给我一个用户记录的表。 日期， 用户名， 活动。 或东是登录。 设计一个发日/周月积极用户的表。 我开始用三个sql 来做。 但即使 扫30 天的数据也很多 。

我使用 一个sql 把3个梳子算出来。 但也需要30 天的书据。 最后他提示设计一个每天用户最后哪天登陆。 最后突然来个里口起吧。

2． 白。 画不同设备（ 手机，电那） 我在这花了一些时间。 要见一个表。

计算人日均铁树。

3． 在大是堂吃饭。 国人很帮忙。 白人爱力不里。

4． 计算电影前x排名。 电影看多少就算。 如百分就十。 里口物流。把看的分钟书加其料。

中见创建多对多关系的表。

我用了窗口函数算排名。

感受

不要淘金湖。 直奔主题。 时间可能不够。 不用解释很多。 你先花。 不多都有提示。

多中被etl 的东西。 注意他们的数据很大。 尽量东最少的书据。 思路一天为单位。

他们要找很多人。 大家多试试。

每轮的full stack都从一个具体的产品出发。例如DAU，update photo，uber。分析用什么metrics衡量产品，如果一些数据drop了，可能是什么原因造成的。之后开始写sql和python。

题目都是这个套路。真的是取决于面试官和面试官的心情决定问什么问题。我碰到的一个面试官问的问题超级发散。感觉不好准备。就靠当时沟通了。其它的几轮还挺常规的。

coding部分lc easy难度。主要考点就是update dict。不用太担心。sql要求还挺高的。当时有一轮还有两个sql没写出来。幸运的过了。self-join, outer join，aggre都是考点。

没有不让写sub query。也没有让写window function。

重点是product sense，我觉得也是bar最高的一部分。lean analysis那本书很经典的，地里也很多人提到过我就不说了。还想给大家推荐一个网站。https://stellarpeers.com/

里面有一些分析的文章很有用！！！重点推荐。还可以和别人约mock。这块重点就是多mock。自己内心有答案真的不代表能表达清楚。我当时基本每块都mock了几遍。问不同人的意见。

也真心感谢帮助我的小伙伴们。分析问题break down的时候也多和面试官沟通。

BQ问题都很常规，conflict，challenge什么的肯定要准备的。还有一个对fb产品的建议。

To test if feature successful:

Define goal

Ask a lot clarifying questions

Define a few metrics to measure the goal

Cohort Analysis: a group of people became user/customer at the same time(group by join date).Track engagement, retention

check for selection bias

self-selection: Specific groups of people may be drawn to taking part in a particular study becauseof self-selecting characteristics.

The best way around for this bias is to draw from a sample that is not self-selecting. This may not always be possible of course, due to experimental

constraints (particularly for studies requiring volunteers), but particular effort should be made to avoid the potential for this bias when examining

different personality types.The effects of this bias are unlikely to be so detrimental if the experiment is concerned with something more constant,

such as psychophysiological measurements.

attrition:

drop out rate is known as participant attrition, and is most commonly seen within investigations where there is an ongoing intervention with several measurements.

If participants drop out of the study in a biased way – if there is a non-random reason why this is occurring – then the remaining participants are unlikely

to be representative of the original sample pool (never mind the population at large).

For example, a medical trial may see numerous participants exit the study if the medicine doesn’t appear to be working (or is making them ill).

In this way, only the remaining (or surviving, in Wald’s case above) participants will be investigated at the end of the experiment.

check for novelty effect. Users use more a product when it is new. Not because it is better, but simply because it is new. As novelty ends, they will use it less.

This is called novelty effect and often makes tests look like winners when they are not.

You can control for this by, in your results, subsetting by drivers for which it is the first experience. Novelty effect obviously doesn't affect new users.

If a test is winning overall, but it is not winning when comparing new users in test vs new users in control, it is a big warning that there might be novelty effect

[more on novelty effect in other questions in this ebook since it is a common topic,often asked like this: "We ran a test. It won by 5%, but, after making the

change for all users and waiting for a couple of weeks, we didn't see any improvement in our metric. Why?" ].

check if you can simply randomly split users, mentally take extreme cases. Let's say the new product has a bug and it is unusable. Or the new product is amazing

and test users will use it 24/7. Will these options have any effect on the control group? If the answer is yes (like it is obvious for the Uber case), you can't

just randomly split users.

Check on cost

Making a change on the site has some costs by itself:

Human labor costs,such as engineering time to make the change, product manager time to document it, customer service time in case some users get confused and

reach out to customer service representatives with questions. Another way to look at this is opportunity-cost. I.e. all these people are not working on

something else with a possibly higher expected value.

Risk of bugs.Whenever you touch the code base, there is a non-zero chance that something will go wrong.

You can't run an A/B test for months. So, whenever your metric needs a long time to be evaluated, you need to find a short term proxy for the long term metric.

This is the point of this question as well as similar questions about long term user engagement, retaining users, or estimating customer lifetime value.

Avoid vanity metrics, i.e. metrics that look good,but are useless. A good way to quickly identify if my metrics are useful is:

imagine a bot is creating a bunch of fake accounts. Will my metrics go up? Inthe example above, we are good. The engagement metric will drop and alert

us that something not good is going on. On the other hand, if you ONLY looked at new sign ups per day, you would be happy with the fake accounts.

If they ask you for just one metric, make sure it is not a metric that a bot could improve.

product growth:

Ability to acquire new users

Ability to retain current users

As you break down the metric and make assumptions along the way, try to end up in a scenario where nothing actually happened on the site, but it simply

changed the user distribution. It makes much easier to give possible reasons as the last step.You can simply take a random country and say users

increased there for a local marketing campaign or decreased because of a local competitor launch. It just makes easier to clearly define and delimitate the problem.

def clac\_time(list):

list.sort(key=lambda x:x[0])

min=0

max=0

total=0

for i in list :

if i[0] &gt; max :

total+=i[1]-i[0]

if i[0] &gt;=min and i[0]max :

total+=i[1]-max

if i[0] max:

max = i[1]

return 'the total is : '+str(total)

FB behavioural questions:

介绍一下你自己.本文原创自1point3acres论坛

最Challenge的project

如何处理同事Conflict

有没有同事或老板给你Critical comments, 之后如何处理

别人东西Delay怎么办. 围观我们@1point 3 acres

Project一开始没方向该怎么办

有External Obstacle怎么处理

最有趣的Project

想去什么组

为什么想加入FB

1. Where do you see yourself in five years?

回答1：I see myself growing into a experienced software engineer. 太泛!

回答2：我有frontend, backend的经验，我希望以后成为一个full stack. 成功安利你懂得不少。

回答3: Become a superviser and use my skills to support others. 显得有team leader的好品质。

另外，根据你申请的公司，如果你知道他们内部升级的机制，可以说两句，比如 Google 的master new grad 一进去是Level3,你可以问问身边的朋友看看一般多久久能升一级。

回答这个问题的时候，就可以带一句你对公司升级机制的了解，希望你2年之内升一级（举例子，我并不知道是不是2年可不可以升啊，不要被我误导）。

来源一亩.三分地论坛.

2. What are your weaknesses?

What are the areas where you need to improve your skills?

Are there areas where you need to develop your skills further?

这类问题叫做negative question。如果回答“I don't have any weakness" 只能说你真不懂面试官意图, 可如果回答 “My greatest weakness is I am working too hard”，这就是在搞事情了。作为一个New grad, 我的weakness就是缺少中作经验，但是答到这还没完。对于这样的negative question，当然要有转折。接下来就要说这不是个问题，我有internship experience, 做过很多projects, tech skills上和team work都有经验，不是个大问题，再来句自己是个fast learner。包括正在申请的同学，如果你想转专业，这个也是个很好的回答方式。如果不是new grad的前辈，可以看看下一题的回答。

关于这个问题的Youtube小视频：https://www.youtube.com/watch?v=4plu0nLyb\_s

3. What do you like least about your previous job/internship?

What do you dislike the most about your last job?

这个就非常尴尬了，是不是觉得这个问题很少问到。我想说的是，我onsite就被问到了这个问题，答的支支吾吾，简直要撕卷。首先，切记！切记！千万不能说之前雇主，同事的问题。即使心里有一万个抱怨，即使是在你看来微不足道的问题，只要是有一点点说 别人的不是，都不是个特别好的回答。没有面试官希望看到雇佣你以后，如果之后你再去面试其他公司，说现在公司和同事的不好。所以这题要怎么答呢。楼主看了个Youtube视频，颇受启发 https://www.youtube.com/watch?v=cK\_KUK1fDc8 以下简介概括。

换个角度，回答这个问题：What do you wish you had done differently at your previous company?

这个问题在未来可以非常容易改正. The problem can be easily changed in the future.

What you have learned from the previous experience.

粗暴的来个例子。

作为了new grad/ intern，希望自己involve到很多的task，所以对这个prioritization做的不是很好。但是知道这是不对的，我现在也慢慢做了调。之前网上找到的，供参考。

I have had problems in the past with taking on too many projects and then not delivering quality and, in some cases, not delivering on time. I was simply stretched too thin in too many areas with not enough time to deliver in a quality manner. I think that's the opportunity that is there as a new college student, you want to do everything and be involved in everything. However（敲黑板）, I've learned that I'm not at my best when I have too many conflicting priorities. I've had to cut back on some of the less important extracurricular activities to focus on delivering my academic projects with the highest quality. This prioritization of my work has carried over to my work life with my recent internship. I had several key deliverables that were due the same week, so I met with my boss to prioritize the delivery schedule. She helped me focus on both the timing and the quality of the delivery. In the end, all three projects were delivered with high quality results.

. 1point 3acres 论坛

Reference: https://collegegrad.com/tough-in ... r-greatest-weakness 这个问题和之前weakness的问题很相似，根据自己的经历，准备好问题都可以变换的回答。

准备Behavior的问题，按照第一段链接里的16个问题，再准备下所谓的negative quesitons就稳了。网上一搜behavior的问题很多很多，但很多都是大同小异。楼主面了无数个面试了，根据自己的经验下面列一下必考题和高频题。

必考题

1. Tell me about yourself

2. Why ..Company?

3. What is your most Challenging project

高频题

1. Describe a situation where you exceeded expectations/did more than required

2. Have you ever had a conflict with the team? How was it resolved? (最好举例)

3. 你想去哪个组，哪个产品，对哪方面感兴趣。当然，如果是面试特定的组，就不会问这个问题。

What do you enjoy about your role?

Answer Question

What (soft skills ) did it take for you to get to where you are?

1 Answer

What do you think it takes to build a great Date Engineering team / Platform

1 Answer

What's the most difficult thing you've had to do?

How to deal with conflict, career plan for the first 90 days.

直奔主题，Onsite一共四场：

Data Modeling 1轮

ETL 2轮

Ownership 1轮

1, 先进行的Ownership，问到的问题有：

Why do you want to work as DE?

Did you ever change your mind because of your teammate?

Did you ever initiate any project/task?

Was there anything that made you unhappy during work?

...