Facebook – Data Scientist

1. Data challenge – ads analysis challenge on the book *the collection of data science take home challenge*?
2. SQL was: you have a table where you have date, user_id, song_id and count. It shows at the end of each day how many times in her history a user has listened to a given song. So count is cumulative sum.
   You have to update this on a daily basis based on a second table that records in real time when a user listens to a given song. Basically, at the end of each day, you go to this second table and pull a count of each user/song combination and then add this count to the first table that has the lifetime count.
   If it is the first time a user has listened to a given song, you won't have this pair in the lifetime table, so you have to create the pair there and then add the count of the last day.
3. Ads related and machine learning questions. How to build an ad model, how to test it, describe a model.
4. Product data analyst role – Facebook products! That's what you'll be analyzing. (Come in with a good understanding of Facebook's line of products in terms of metrics, features, audience.)
5. How would you measure the health of Mentions, Facebook's app for celebrities? How can FB determine if it's worth it to keep using it?
   If a celebrity starts to use Mentions and begins interacting with their fans more, what part of the increase can be attributed to a celebrity using Mentions, and what part is just a celebrity wanting to get more involved in fan engagement?
6. There is a table that tracks every time a user turns a feature on or off, with columns user_id, action ("on" or "off"), date, and time.
   How many users turned the feature on today?
   How many users have ever turned the feature on?
   In a table that tracks the status of every user every day, how would you add today's data to it?
7. If 70% of Facebook users on iOS use Instagram, but only 35% of Facebook users on Android use Instagram, how would you investigate the discrepancy?
8. How do you measure newsfeed health?
9. If a PM says that they want to double the number of ads in Newsfeed, how would you figure out if this is a good idea or not?
10. We have two options for serving ads within Newsfeed:
    1 - out of every 25 stories, one will be an ad
    2 - every story has a 4% chance of being an ad
    For each option, what is the expected number of ads shown in 100 news stories?
    If we go with option 2, what is the chance a user will be shown only a single ad in 100 stories? What about no ads at all?
11. How do you map nicknames (Pete, Andy, Nick, Rob, etc) to real names?
12. Facebook sees that likes are up 10% year over year, why could this be?
13. How many high schools that people have listed on their profiles are real? How do we find out, and deploy at scale, a way of finding invalid schools? (mapping users' high school) How do you justify whether Facebook users' high school information is correct?
14. Product questions that ask you how to derive a specific metric or what data you would use to solve a specific problem. Pages, Likes, Ads.
15. Statistics, conditional probability, dice/deck of cards, distribution. Expectation for a probability distribution, e.g.

geometric distribution

16. How to use T-test to compare the results of different experiments
17. Why linear regression is not suitable for advanced classification task (has a lot of assumptions?)
18. A table has a column with full names (e.g. 'Tom Hardin'). write a python code to get distinct first names(i know its super easy in sql)
19. How do you measure user engagement, how do you measure the effectiveness of a new feature (metrics?). Thought process.
20. How do you calculate monthly active users, churned users and resurrected users from a user activity log with userID and DateTime
21. What aspects of available data you would use to answer certain business questions or anomalies
22. Say you notice a surge in average activity among users but the number of users is down, how would you explain this?
23. Estimate why would the ratio of like/daily active users would increase suddenly?
24. Compute the distribution of likes per post
25. Given a list, create a new list that does not include duplicate values of the original list.
26. How would you find Facebook user's best friend?
27. Reviewing a popular app?
28. They will ask you to design a very simple thing. Be prepared to innovate on the spot.
29. Flip one coin 10 times and obtain one H, what is the p value and null hypothesis?
30. Suppose the Recommended Friends team has updated the algorithm to analyze if the new algorithm is better than the old algorithm.
31. Given the following tables how would you know who has the most friends
    REQUESTS
    date | sender_id | accepter_id
    ACCEPTED
    accepted_at | accepter_id | sender_id
32. Black jack probably from card decks
33. If there's a new feature that attracts more pop stars to use it, how do you evaluate whether this feature is successful or not?
34. Given tables:
    employees(id, unixname,team,role,days_since_started)
    projects(id,name,....)
    commits(id,file_path,proj_id,auth_id,timestamp)
    Find the number of unique employees per project per month?
35. Imagine you have a CSV file: (python question?)
    john_doe,android,ios,infra
    bob_law,is,backend
    jane_doe,frontend

    Containing, for each employee, a list of projects he/she works on
    Write a script/function/else that reads data from CSV file and creates a data structure that stores, for each project,

a list of employees who work on it (people who are assigned to it).

36. How would you convince someone that two coins being tossed are both fair?
37. How do you proof that males are on average taller than females by knowing just gender or height
38. How do you get the count of each letter in a sentence (python?)
39. Calculate various statistics from two tables: one for advertisers, and one for ads
40. How would you explain a confidence interval to a non-technical audience?
41. Tell me about a time when you received constructive criticism and how you handled it.
42. Define KPIs (metrics?)
43. Given an event-level table of interactions between pairs of users (note that there aren't duplicates in one day for one pair of users), for each possible number of "people interacted with" find the count for that group in a given day (i.e. 10 people interacted with only one person, 20 with 2, etc.).
44. A product manager is considering replacing an ad unit with a 'people you may know' unit on the news feed. What's your advice?
45. Bias-variance trade-off for linear regression versus nonlinear fits
46. product questions about what could be driving activity, what could be causing this, what data would you look at
47. You have a table with appID, eventID, and timestamp. eventID is either 'click' or 'impression'. Calculate the click through rate. Now do it in for each app.
48. Likes/user and minutes spent on a platform are increasing but total number of users is decreasing. What could be causing this?
49. Then he just throw out some product-related questions, with a bunch of follow up questions. Remember to be well prepared for what you are saying, cuz the interviewer may ask you why for some details, and you need also to make your thoughts well organized, since multiple reasons are required (I think they expect the insights shown in your answer, not the generalized universal truth).
50. How to determine if a conservation was happening in the comments algorithmically
51. How can we tell if two users on Instagram are best friends?
    (Answer question). Ok, how can we use this algorithm in the product?
52. Industry terminology. Countermetrics?
53. What is data science?
54. How would you tell whether a new feature X is beneficial or not
55. There is a feature for showing friend requests. How do you measure its impact and how do you decide whether to keep it.
56. If revenue is down 1% and time spent is up 2%, what do you infer/decide?
57. A set of values given: Assume table in SQL or list of dictionaries if using Python.
    Basically a row of data contained information:
    if it is post or it is a comment, row id and some other data.
    Find distribution of comments.
    #comments # posts
    1 5000
    2 6787
    ..
58. How would you measure success of a new product?

59. SQL assignment was to write a query which computes conversion rates for 2 products given the log of user actions with type of action for each product type.
60. What is your favorite Facebook product?
61. Write a sql query to find out the overall friend acceptance rate for a given date?
    Table : User_id_who_sent|User_id_to_whom|date|Action (Sent, accepted, rejected etc)
62. What would be your approach to check if the overall acceptance rate has decreased from 60% in may'17 to 30% in sep'17
63. How would you conduct an experiment to test if a change in facebook app is effective and what metrics will you look at?
64. What metric would you show small businesses if you were trying to have them sign up for Facebook Ads
65. Given a table of friend requests sent and friend requests received, find the user with the most friends.
66. If you implement a new feature to FB how would you measure the success
67. Question about evaluating satisfaction of a product.
68. Developing a metric, how to setup an experiment, making decision based on result
69. Discuss pros and cons of using like-rate to determine system health
70. Discuss what might cause average like rate to go from 70% one day to 30% another
71. Explain how you would diagnose a 10% YoY "like" improvement
72. What would be your next step to investigating why this happened
73. How would you improve the product/service?
74. Make sure you are familiar with the differences between Facebook/Instagram/Snapchat/Twitter in terms of social graphs, UX, monetization etc.
75. Make a histogram of 2 variables. Building a histogram of post reply count in SQL (number of posts with x replies, x+1 replies, etc).
76. Building models to solve specific FB problems (are people connected/not), identifying pitfalls in a model and data leakage?
77. A/B testing (if we have a product and see some trend in different users make a hypothesis about why we see that trend and test it)
78. SWE with a CS type algorithm development problem?
79. Full disclosure I do not have/use Facebook or Instagram and though I disclosed this before beginning the process I felt that it had an impact on my interviews.
80. We have a product that is getting used differently by two different groups. What is your hypothesis about why and how would you go about testing it?
81. Given a specific product, come up with some potential improvements and design a series of experiments for testing/implementing these changes.
82. The company developed a new feature and perform A/B test. Here is the result
    Comments +5%
    Likes -10%
    Timespent +1%
    All else neutral
    How would you decide to whether putting into product based on the A/B test result? Any ideas?
83. What is a Facebook product that you'd be interested in working on, and how would you use analytics?

84. We've introduced this particular new feature (it was something about friend recommendations). How should we determine whether to keep it or not?
85. SQL was a straightforward problem that involved calculating the friend acceptance rate over time. The table consisted of events (friend request or acceptance), date, and user ids.
86. Building a table with a summary of feature usage per user every day (keep track of the last action by user and roll that up every day)
87. Give me an analytics project you have done?
88. Imagine we see a lot of users filling up a form but not submitting it, why would it be the case and how would you use data to finding it out?
89. Online coding question using mysql about like/dislike system in facebook
90. How to choose emotions in like/dislike systems and any problems in the A/B test.
91. What are the assumptions of OLS-Regression?
92. Given a list A of objects and another list B which is identical to A except that one element is removed, find that removed element.
93. What experiment would you run to implement new features on Facebook?
94. How would you test the success of a new feature added to a product (e.g. Messenger) and what variables to explore
95. business insight question (pros and cons of a method, they seems to be interested in revenue aspect
96. The success ratio of sending messages given sent and receive table
97. A product/business sense section where you discuss FB products, how you would improve them, and what type of metrics you would use to gauge success
98. What is the expectation of the variance?
99. There's a game where you are given two fair six-sided dice and asked to roll. If the sum of the values on the dice equals seven, then you win $21. However, you must pay $5 to play each time you roll both dice. Do you play this game?
    And in follow-up:
    What is the probability of making money from this game?
100. We at Facebook would like to develop a way to estimate the month and day of people's birthdays, regardless of whether people give us that information directly. What methods would you propose, and data would you use, to help with that task?
101. There are two mobile restroom stalls at a construction site where I work.
    There are also three situations that have an equal chance of occurrence:
    a. none of them is occupied
    b. only one of them is occupied
    c. both are occupied
    1. If I were to pick one at random, what is the probability that it is occupied?
    2. If it turns out that that first one I go to is occupied and I decide to try the other one, what is the probability that the second one is also occupied?
102. Why the number of likes are increasing?
103. Define the right metric for certain FB products
104. Business problem: Implementing creation of new reaction like happy, sad etc.
105. How could graph theory be used to verify users' information?

106. How can Facebook figure out when users falsify their attended schools?
107. frequentist inference
108. If we were to add X feature to Facebook, how would you measure if it was a good idea or not? A discussion ensued with each step leading to the next.
109. In Mexico, if you take the mean and the median age, which one will be higher and why?
110. If you draw 2 cards from a shuffled 52 card deck, what is the probability that you'll have a pair?
111. Given a table that each day shows who was active in the system and a table that tracks ongoing user status, write a procedure that will take each day's active table and pass it into the ongoing daily tracking table.
    Possible states are:
    * user stayed (yesterday yes, today yes)
    * user churned (yesterday yes, today no)
    * user revived (yesterday no, today yes)
    * user new (yesterday null, today yes)
    Note: you'll want to spot and account for the undefined state.
112. There are 50 cards of 5 different colors. It comprises of 10 Red cards, 10 blue cards, 10 orange cards, 10 green cards and 10 yellow cards. Each color will have the cards numbered between 1 to10. You pick 2 cards at random. What is the probability that they are not of same color and not of same number.
113. What is one project you're proud of
114. defining user engagement metric
115. What is the probability of drawing two cards (from the same deck of cards) that have the same suite?
116. Find the expected revenue/loss of throwing two dices.
117. Interviewed me about a hypothetical feature -- a credit card info pop-up?
118. Product development question -- how would one develop the Like, Love, Sad feature
119. What is probability of getting one pair of card from 52 deck of cards?
120. How to evaluate the success of a product?
121. Programming: Given data on Facebook members friending/defriending each other on Facebook, find out whether a given pair of members are currently friends
122. Statistics: Very basic questions on treatment effects under various sampling schemes
123. Research Design Question: How would you test whether having more friends now increases the probability that a Facebook member is still an active user after 6 months?
124. Given two tables
    Friend_request
    (requester_id,
    sent_to_id,
    time)
    Request_accepted
    (acceptor_id,
    requestor_id,
    time)
    Find the overall acceptance rate of requests.
125. Open ended questions (i.e. how would you do this? what kind of problems do you think you will encounter with

this approach? etc.). After you would describe your approach, they would go into details a lot to make sure you actually have a clear idea of how to take the project from start to end

126. How have you dealt with a difficult person in your current job?
127. What do you not like about your current job? (Idea hard to reach implementation phase?)
128. Coding on whiteboard: solutions required using recursion and hash tables?
129. How to communicate statistical analysis to non-technical managers
130. Basic statistics: solving a problem with a two sample t-test, dealing with multicollinearity
131. Standard questions on machine learning: cross-validation, regularization methods?
132. Write an SQL query that makes recommendations using the pages that your friends liked. Assume you have two tables: a two-column table of users and their friends, and a two-column table of users and the pages they liked. It should not recommend pages you already like.
133. How to generate all possible friendship given four tables with the "request", "accept","reject", and "remove" information.
134. Why Facebook?
135. What is probability of pulling a different color or shape card from a shuttled deck of 52 cards?
136. How would you compare the relative performance of two different back-end engines for automated generation of Facebook "Friend" suggestions?
137. Did a problem in SQL and a followed question about product – about what can I do given the table in SQL and some other variables added
138. How to detect the fake information input from the user
139. How many Big Mac does McDonald sell each year in US?
140. You a 2 dices. What's the probability of getting at least one 4? Same question with n dice.
141. How many dentist are there in the US?
142. Answer some strategic questions about how you might think about the SQL output or what conclusions you could draw from the metrics available to you (or those you might create).
143. What kind of metrics I would look at, how I would track them, what conclusions I might draw from different scenarios
144. Probability – Bayes rule?
145. What is standard deviation?
146. You're at a casino with two dice, if you roll a 5 you win, and get paid $10. What is your expected payout? If you play until you win (however long that takes) then stop, what is your expected payout?
147. Three ants are sitting at the three corners of an equilateral triangle. Each ant starts randomly picks a direction and starts to move along the edge of the triangle. The probability that none of the ants collide = [? / ?] Follow-up: k ants are sitting at the k corners of an equilateral polygon. Each ant starts randomly picks a direction and starts to move along the edge of the triangle. The probability that none of the ants collide = [? / ?]
    Count how many trailing 0 in (100!)
148. If we were testing product X, what metrics would you look at to determine if it is a success?
149. Consider a game with 2 players, A and B. Player A has 8 stones, player B has 6. Game proceeds as follows. First, A rolls a fair 6-sided die, and the number on the die determines how many stones A takes over from B. Next, B rolls the same die, and the exact same thing happens in reverse. This concludes the round. Whoever has more stones at the end of the round wins and the game is over. If players end up with equal # of stones at the end of the round,

it is a tie and another round ensues. What is the probability that B wins in 1, 2, ..., n rounds?

150. Lets say the population on Facebook clicks ads with a click-through-rate of P. We select a sample of size N and examine the sample's conversion rate, denoted by hat{P}, what is the minimum sample size N such that Probability(ABS(hat{P} - P) < DELTA) = 95%. In other words (this is my translation), find the minimum sample size N such that our sample estimate hat{P} is within DELTA of the true click through rate P, with 95% confidence.
151. Write a SQL query to compute a frequency table of a certain attribute involving two joins. What if you want to GROUP or ORDER BY some attribute? What changes would you need to make? How would you account for NULLs?
152. What teams are you interested in at Facebook?
153. Pick up a coin C1 given C1+C2 with probability of trials p(h1)=.7, p(h2)=.6 and doing 10 trials. And what is the probability that the given coin you picked is C1 given you have 7 heads and 3 tails?
154. What metrics you would look at for a particular product, how would you interpret them, etc.
155. There are two types of cars A and B. The number of people in US who use A and B are the same. They drive the same distances each month.

    Now there are two new technologies, X and Y (of equal cost).

    If apply X, mpg of A would increase from 50 mpg to 75 mpg;

    If apply Y, mpg of B would increase from 10 mpg to 11 mpg.

    The goal is to decrease the dependence on foreign oil, or to decrease the consumption of gasoline.

    Question: which technology would you apply?

    Follow up question: after applying the technology of your choice, assume there's money available for research on new technology, which car would you choose to conduct research on?
156. dialoglog

    (userid int

    appid int

    type char , a flag either "imp" or "click"

    ds timestamp)

    How would you access the quality of app?

    How to compute click-through rate (in mySQL)?
157. Given a list, create a new list that does not include the duplicates of the original list.
158. Calculate yearly tips for taxi manager in LA
159. If X1 and X2 are normally distributed (and IID) with mu, sigma, and Y=X1+X2, what is the mean and s-dev of Y?
160. You're about to get on a plane to Seattle. You want to know if you should bring an umbrella. You call 3 random friends of yours who live there and ask each independently if it's raining. Each of your friends has a 2/3 chance of telling you the truth and a 1/3 chance of messing with you by lying. All 3 friends tell you that "Yes" it is raining. What is the probability that it's actually raining in Seattle?
161. What would you add to Facebook and how would you pitch it and measure its success.
162. Write a function that takes in two sorted lists and outputs a sorted list that is their union.
163. They spend time figuring out how to run a test to increase $$ from subscription without waiting 6 months to analyze the results. Or how to split users in two independent groups for an A/B test if the site is a social network and users are all connected. Or like in that question from the book, how do you test a new referral program? Try to go through it step by step in your mind, how would you do it? As you think through it, you will realize that there are many problems. Think about how to solve them. Very often when I interview people I feel they never deeply

thought about these problems before the interview. And that's why they fail.

164. Do a deep dive on the company -- What data do they have, how is it stored, and what kind of complications might they run into? What are their key performance indicators, and what kind of projects would they want to use their data for? How would you implement those projects?
165. Going through the technical blog of the company will give you enough jargon to talk your way out of such questions. e.g. https://eng.uber.com/
166. The interviewer will ask questions specific to the business. A typical question could be – How would you improve [main metric that team cares about]? It could also be how would you improve mobile usage, response time, subscription retention, churn or search depending on the team you're interviewing with. This usually boils down to describing how you would approach the problem, do the analysis, give examples of a segment that performs badly, explain why you think that segment is doing poorly and come up with an idea to fix that. It is essential you have a good understanding of the business, top products, how the business makes money and risk factors. You should also be prepared for open ended / case questions since it is the core of the onsite interview at a lot of places where Data Science sits within the product team.