What Top Firms Ask: 100+ Data Science Interview Questions

Previous post
Next post

Tags: Algorithms, Data Science, Google, Hadoop, Interview questions, Machine Learning, Microsoft, Statistics, Uber

Check this out: A topic wise collection of 100+ data science interview questions from top companies.

A fresh scrape from Glassdoor gives us a good idea about what applicants are asked during a data scientist interview at some of the top companies. Unfortunately for us, almost every company has their interviewees sign NDAs. Since Glassdoor allows anonymity, a few brave souls have given us some fantastic examples of what they were asked during the interview process at top companies like Facebook, Google, and Microsoft.

If you find yourself unable to answer some of the questions below, consider checking out a course or a book on the subject.

If you'd like to share your answer(s) to any of the questions, leave a comment and I'll add the top ones to the post. Just make sure to comment with your real name so I can give you credit!

Also, if you don't see a particular question on this list that you've been asked, or you know of one that's asked a lot, comment below. I'd love to add it.

General Questions:

Apple

1. Suppose you're given millions of users that each have hundreds of transactions and these millions of transactions are for tens of thousands of products. How would you group the users together in meaningful segments?

Microsoft

2. Describe a project you've worked on and how it made a difference.
3. How would you approach a categorical feature with high-cardinality?
4. What would you do to summarize a Twitter feed?
5. What are the steps for wrangling and cleaning data before applying machine learning algorithms?
6. How do you measure distance between data points?
7. Define variance.
8. Describe the differences between and use cases for box plots and histograms.

Twitter

9. What features would you use to build a recommendation algorithm for users?

Uber

10. Pick any product or app that you really like and describe how you would improve it.
11. How would you find an anomaly in a distribution ?
12. How would you go about investigating if a certain trend in a distribution is due to an anomaly?
13. How would you estimate the impact Uber has on traffic and driving conditions?
14. What metrics would you consider using to track if Uber's paid advertising strategy to acquire new customers actually works? How would you then approach figuring out an ideal customer acquisition cost?

LinkedIn

15. Big Data Engineer Can you explain what REST is?

Machine Learning Questions:

Google

16. Why do you use feature selection?
17. What is the effect on the coefficients of logistic regression if two predictors are highly correlated? What are the confidence intervals of the coefficients?
18. What's the difference between Gaussian Mixture Model and K-Means?
19. How do you pick k for K-Means?
20. How do you know when Gaussian Mixture Model is applicable?
21. Assuming a clustering model's labels are known, how do you evaluate the performance of the model?

Microsoft

23. Choose any machine learning algorithm and describe it.
24. Describe how Gradient Boosting works.
25. Data Mining Describe the decision tree model.
26. Data Mining What is a neural network?
27. Explain the Bias-Variance Tradeoff
28. How do you deal with unbalanced binary classification?
29. What's the difference between L1 and L2 regularization?

Uber

30. What sort features could you give an Uber driver to predict if they will accept a ride request or not? What supervised learning algorithm would you use to solve the problem and how would compare the results of the algorithm?

LinkedIn

31. Name and describe three different kernel functions and in what situation you would use each.
32. Describe a method used in machine learning.
33. How do you deal with sparse data?

IBM

34. How do you prevent overfitting?
35. How do you deal with outliers in your data?
36. How do you analyze the performance of the predictions generated by regression models versus classification models?
37. How do you assess logistic regression versus simple linear regression models?
38. What's the difference between supervised learning and unsupervised learning?
39. What is cross-validation and why would you use it?
40. What's the name of the matrix used to evaluate predictive models?
41. What relationships exist between a logistic regression's coefficient and the Odds Ratio?
42. What's the relationship between Principal Component Analysis (PCA) and Linear & Quadratic Discriminant Analysis (LDA & QDA)
43. If you had a categorical dependent variable and a mixture of categorical and continuous independent variables, what algorithms, methods, or tools would you use for analysis?

Salesforce

45. What data and models would would you use to measure attrition/churn? How would you measure the performance of your models?
46. Explain a machine learning algorithm as if you're talking to a non-technical person.

Capital One

47. How would you build a model to predict credit card fraud?
48. How do you handle missing or bad data?
49. How would you derive new features from features that already exist?
50. If you're attempting to predict a customer's gender, and you only have 100 data points, what problems could arise?
51. Suppose you were given two years of transaction history. What features would you use to predict credit risk?
52. Design an AI program for Tic-tac-toe

Zillow

53. Explain overfitting and what steps you can take to prevent it.
54. Why does SVM need to maximize the margin between support vectors?

Hadoop:

Twitter

55. How would you use Map/Reduce to split a very large graph into smaller pieces and parallelize the computation of edges according to the fast/dynamic change of data?
56. Data Engineer Given a list of followers in the format:123, 345234, 678345, 123…Where column one is the ID of the follower and column two is the ID of the followee. Find all mutual following pairs (the pair 123, 345 in the example above). How would you use Map/Reduce to solve the problem when the list does not fit in memory?

Capital One

57. Data Engineer What is Hadoop serialization?
58. Explain a simple Map/Reduce problem.

Hive:

LinkedIn

Spark:

Capital One

60. Data Engineer Explain how RDDs work with Scala in Spark

Statistics & Probability Questions:

Google

61. Explain Cross-validation as if you're talking to a non-technical person.
62. Describe a non-normal probability distribution and how to apply it.

Microsoft

63. Data Mining Explain what heteroskedasticity is and how to solve it

Twitter

64. Given Twitter user data, how would you measure engagement?

Uber

65. What are some different Time Series forecasting techniques?
66. Explain Principle Component Analysis (PCA) and equations PCA uses.
67. How do you solve Multicollinearity?
68. Analyst Write an equation that would optimize the ad spend between Twitter and Facebook.

Facebook

69. What's the probability you'll draw two cards of the same suite from a single deck?

IBM

70. What are p-values and confidence intervals?

Capital One

71. Data Analyst If you have 70 red marbles, and the ratio of green to red marbles is 2 to 7, how many green marbles are there?
72. What would the distribution of daily commutes in New York City look like?
73. Given a die, would it be more likely to get a single 6 in six rolls, at least two 6s in twelve rolls, or at least one-hundred 6s in six-hundred rolls?

74. What's the Central Limit Theorem, and how do you prove it? What are its applications?

Programming & Algorithms:

Google

75. Data Analyst Write a program that can determine the height of an arbitrary binary tree

Microsoft

76. Create a function that checks if a word is a palindrome.

Twitter

77. Build a power set.

78. How do you find the median of a very large dataset?

Uber

79. Data Engineer Code a function that calculates the square root (2-point precision) of a given number. Follow up: Avoid redundant calculations by now optimizing your function with a caching mechanism.

Facebook

80. Suppose you're given two binary strings, write a function adds them together without using any builtin string-to-int conversion or parsing tools. For example, if you give your function binary strings 100 and 111, it should return 1011. What's the space and time complexity of your solution?

81. Write a function that accepts two already sorted lists and returns their union in a sorted list.

LinkedIn

82. Data Engineer Write some code that will determine if brackets in a string are balanced

83. How do you find the second largest element in a Binary Search Tree?

84. Write a function that takes two sorted vectors and returns a single sorted vector.

85. If you have an incoming stream of numbers, how would you find the most frequent numbers on-the-fly?

86. Write a function that raises one number to another number, i.e. the pow() function.

87. Split a large string into valid words and store them in a dictionary. If the string cannot be split, return false. What's your solution's complexity?

Salesforce

88. What's the computational complexity of finding a document's most frequently used words?

Capital One

90. Data Engineer How would you ‘disjoin’ two arrays (like JOIN for SQL, but the opposite)?
91. Create a function that does addition where the numbers are represented as two linked lists.
92. Create a function that calculates matrix sums.
93. How would you use Python to read a very large tab-delimited file of numbers to count the frequency of each number?

PayPal

94. Write a function that takes a sentence and prints out the same sentence with each word backwards in O(n) time.
95. Write a function that takes an array, splits the array into every possible set of two arrays, and prints out the max differences between the two array’s minima in O(n) time.
96. Write a program that does merge sort.

SQL Questions:

Microsoft

97. Data Analyst Define and explain the differences between clustered and non-clustered indexes.
98. Data Analyst What are the different ways to return the rowcount of a table?

Facebook

99. Data Engineer If you’re given a raw data table, how would perform ETL (Extract, Transform, Load) with SQL to obtain the data in a desired format?
100. How would you write a SQL query to compute a frequency table of a certain attribute involving two joins? What changes would you need to make if you want to ORDER BY or GROUP BY some attribute? What would you do to account for NULLS?

LinkedIn

101. Data Engineer How would you improve ETL (Extract, Transform, Load) throughput?

Brain Teasers & Word Problems:

Google

102. Suppose you have ten bags of marbles with ten marbles in each bag. If one bag weighs differently than the other bags, and you could only perform a single weighing, how would you figure out which one is different?

103. You are about to hop on a plane to Seattle and want to know if you should carry an umbrella. You call three friends of yours that live in Seattle and ask each, independently, if it's raining.

104. Each of your friends will tell you the truth ⅔ of the time and mess with you by lying ⅓ of the time. If all three friends answer "Yes, it's raining," what is the probability that is it actually raining in Seattle?

Uber

105. Imagine you are working with a hospital. Patients arrive at the hospital in a Poisson Distribution, and the doctors attend to the patients in a Uniform Distribution. Write a function or code block that outputs the patient's average wait time and total number of patients that are attended to by doctors on a random day.

Facebook

106. Imagine there are three ants in each corner of an equilateral triangle, and each ant randomly picks a direction and starts traversing the edge of the triangle. What's the probability that none of the ants collide? What about if there are N ants sitting in N corners of an equilateral polygon?

107. How many trailing zeros are in 100 factorial (i.e. 100!)?

LinkedIn

108. Imagine you're climbing a staircase that contains n stairs, and you can take any number k steps. How many distinct ways can you reach the top of the staircase? (This is a modification of the original stair step problem)

Questions sourced from Glassdoor

Original. Reposted with permission.

Related:

- 7 More Must-Know Data Science Interview Questions and Answers
- 17 More Must-Know Data Science Interview Questions and Answers, Part 2
- 17 More Must-Know Data Science Interview Questions and Answers, Part 3

Previous post
Next post

Top Stories Past 30 Days

Most Popular

1. 30 Essential Data Science, Machine Learning & Deep Learning Cheat Sheets
2. Understanding Machine Learning Algorithms
3. Introduction to Blockchains & What It Means to Big Data
4. Want to Become a Data Scientist? Read This Interview First
5. The 10 Algorithms Machine Learning Engineers Need to Know
6. Keras Cheat Sheet: Deep Learning in Python
7. How I started with learning AI in the last 2 months

Most Shared

1. Want to Become a Data Scientist? Read This Interview First
2. Using Machine Learning to Predict and Explain Employee Attrition
3. Top 10 Machine Learning Algorithms for Beginners
4. An Overview of 3 Popular Courses on Deep Learning
5. How LinkedIn Makes Personalized Recommendations via Photon-ML Machine Learning tool
6. Deep Learning for Object Detection: A Comprehensive Review
7. [webinar] Getting Started with Automated Analytics Powered By

Latest News

- Recommendation Engines and Real-time personal...
- Spotify Global VP Opens Data Marketing Toront...
- The danger in comparing your campaign perform...
- Hello, World: Building an AI that understands...
- Density Based Spatial Clustering of Applicati...
- Arena: Sr. Data Scientist

More Recent Stories

- Arena: Sr. Data Scientist
- Top tweets, Oct 18-24: Chihuahua or muffin? The #DataScienc...
- AI Expo North America, Santa Clara, Nov 29-30, 2017
- Xavier U. of Louisiana: Assistant Professor, Data Science
- Neural Network Foundations, Explained: Updating Weights with G...
- Build, Test and Run Spark Applications at No Cost with Stream...
- Artificial Intelligence Today: Time to Act
- No order left behind; no shopper left idle.
- The Humanalysts: Data Analytics Team, 6
- KDnuggets 17:n41, Oct 25: Learning git not enough to become...
- Domino Data Science Pop-up – Chicago, Nov 14
- Applied AI Summit will give you the tools for your AI journey,...
- Top 10 Machine Learning with R Videos
- Business intuition in data science
- RWTH Aachen University: 18 Process Mining Jobs (11 PhDs, 5 Pos...
- How Can Machine Learning Affect Your Organizational Data Strat...
- Institute for Defense Analyses: Jr. Data Scientist
- Kanri Distance Calculator Free License Version with Demo
- Your Complete Guide to Predictive Analytics World – Oct ...
- Ranking Popular Deep Learning Libraries for Data Science

X