BitDegree

DATA SCIENCE

INTERVIEW QUESTIONS LEAKED

Learn What Will Get You Hired

INTRODUCTION

Naturally, there's a huge need for qualified data scientists in the market. The job opportunities for this position are constantly increasing. So if you're thinking about applying for a data scientist job position, you'll need to know the essential data science interview questions. This tutorial will provide you with exactly that.

This book is split into two big parts - the basics and the more advanced stuff. Well talk about big data interview questions, differentiate data scientists from data analysts and so on. At the very end, I'll give you a couple of tips to stay cool during your interviews and what people that have worked thousands of hours in the industry expect from potential employers.

A lot of your early data science interview questions might include differentiating between seemingly similar, yet somewhat different terms. That's why it's probably a good idea to start from these definitions so that you have a clear understanding of what is what moving forward.

Common **Interview** Questions

1 What is 'Data Science'?

Data science is a form of methodology that is used to extract and organize various data and information out of huge data sources (both structured and unstructured).

The way that this form of science works is that it uses various algorithms and applied mathematics to extract useful knowledge and information and arrange it in a way that would make sense and grant some sort of usage.

2 Big Data Vs. Data Science

Surely one of the more tricky data science interview questions, a lot of people fail to express a clear difference. This is mostly because of a lack of information surrounding the topic.

However, the answer itself is actually very simple - since the term 'big data' implies huge volumes of data and information, it needs a specific method to be analyzed. So, **big data is the thing that data science analyzes.**

3 Leaked Interview Assignment

What's the difference between a 'data scientist' and a 'data analyst'?

Even though this is also one of the basic data science interview questions, the terms still often tend to get mixed up.

Data scientists mine, process and analyze data. They are concerned with providing predictions for businesses on what problems they might come across.

Data analysts solve the unavoided business problems instead of predicting them beforehand. They identify issues, perform analysis of statistical information and document everything.

4 The Core Features of Big Data

Now that we've covered the definitions, we can move to the specific data science interview questions. Keep in mind, though, that you are bound to receive data scientist, analyst and big data interview questions. The reason why is because all of these subcategories are intertwined with each other.

There are five categories that represent big data, and they're called the "5 Vs":

1. Value
2. Variety
3. Velocity
4. Veracity
5. Volume

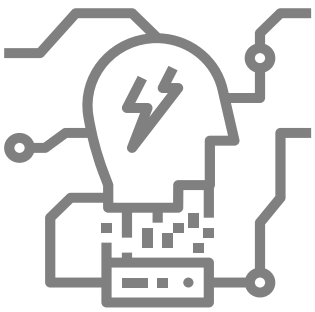
5 What's a 'Recommender System'?

It is a type of system that is used for predicting how high of a rating would users give to certain specific objects (movies, music, merchandise, etc.). Needless to say, there are a lot of complex formulas involved in such a system.

6 What's a 'Power Analysis'?

A type of analysis that's used to determine what sort of an effect will a unit have simply based on its size.

Power analysis is directly related to tests of hypotheses. The main purpose underlying power analysis is to help the researcher to determine the smallest sample size that is suitable to detect the effect of a given test at the desired level of significance.



7 What's A/B Testing?

While A/B testing can be applied in various different niches, it is also one of the more prominent data science interview questions. So what is it?

A/B testing is a form of tests conducted to find out which version of the same thing is more worth using to achieve the desired result.

Say, for example, that you want to sell apples. You're not sure what type of apples - red or green ones - your customers will prefer. So you try both - first you try to sell the red apples, then the green ones. After you're done, you simply calculate which were the more profitable ones and that's it - that's A/B testing!

8 What's 'Hadoop'?

Hadoop is an open source distributed processing framework that manages data processing and storage for big data applications running in clustered systems.

Apache Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model.

Hadoop splits files into large blocks and distributes them across nodes in a cluster. It then transfers packaged code into nodes to process the data in parallel. This allows the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture.

9 What's a 'Selection Bias'?

Selection bias is the bias introduced by the selection of individuals, groups or data for analysis in such a way that proper randomization is not achieved, thereby ensuring that the sample obtained is not representative of the population intended to be analyzed.

If the selection bias is not taken into account, then some conclusions of the study may not be accurate.

10 Define 'Collaborative Filtering'?

Collaborative filtering, as the name implies, is a filtering process that a lot of recommender systems utilize. This type of filtering is used to find and categorize certain patterns.

Collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). This type of filtering is used to find and categorize certain patterns.

11 What's 'fsck'?

'fsck' abbreviates as "File System Check". It is a type of command that looks for possible errors within the file and, if there are errors or problems found, fsck reports them to the Hadoop Distributed File System.

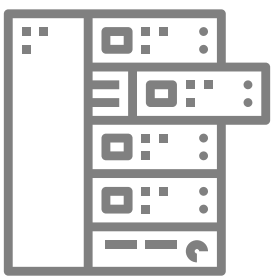
12 What's a 'Cross-validation'?

Yet another addition to the data analyst interview questions, cross-validation can be quite difficult to explain, especially in a simplistic and easily understandable manner.

Cross-validation is used to analyze if an object will perform the way that it is expected to perform once put on the live servers. In other words, it checks how certain results of specific statistical analyses will measure when placed into an independent set of data.

13 What's 'Cluster Sampling'?

Cluster sampling refers to a type of sampling method. With cluster sampling, the researcher divides the population into separate groups, called clusters. Then, a simple random sample of clusters is selected from the population. The researcher conducts his analysis on data from the sampled clusters.



Advanced **Interview** Questions

14 Bonus: Possible Interview Exercise

Which is better - good data or good models?

The answer to this question is truly very subjective and case-by-case dependant. Bigger companies might prefer good data, for it is the core of any successful business. On the other hand, good models couldn't really be created without having good data.

You should probably pick according to your own personal preference - there really isn't any right or wrong answer (unless the company is specifically searching for either one of them). So, do your research about the company. Try to see if they're testing your knowledge of their product or is it a 'trick question'.

15 Bonus: Possible Interview Exercise 2

What's the difference between 'supervised' and 'unsupervised' learning?

Although this isn't one of the most common data scientist interview questions and has more to do with machine learning than with anything else, it still falls under the umbrella of data science, so it's worth knowing.

During supervised learning, you would infer a function from a labeled portion of data that's designed for training. Basically, the machine would learn from objective and concrete examples that you provide.

Unsupervised learning refers to a machine training method which uses no labeled responses - the machine learns by descriptions of the input data.

16 'Expected Value' Vs. 'Mean Value'?

When it comes to functionality, there's no difference between the two. However, they are both used in different situations.

Expected values usually reflect random variables, while mean values reflect the sample population.

17 'Bivariate' Vs. 'Multivariate' and 'Univariate'

A bivariate analysis is concerned with two variables at a time, while multivariate analysis deals with multiple variables. Univariate analysis is the simplest form of analyzing data. "Uni" means "one", so in other words, your data has only one variable. It doesn't deal with causes or relationships (unlike regression) and its major purpose is to describe; it takes data, summarizes that data and finds patterns in the data.

18 Bonus: Possible Interview Exercise 3

What if two users were to access the same HDFS file at the same time?

This is also one of the more popular data scientist interview questions - and it's somewhat of a tricky one. The answer itself isn't difficult at all, but it's easy to mix it up with how similar programs react. If two users are trying to access a file in HDFS, the first person gets the access, while the second user (that was a bit late) gets denied.

How many common Hadoop input formats are there? What are they?

One of the interview questions for data analyst that might also show up in the list of data science interview questions. It's difficult because you not only need to know the number, but also the formats themselves.

In total, there are three common Hadoop input formats. They go as follows: key-value format, sequence file format and text format.

19 Bonus: Possible Interview Exercise 4

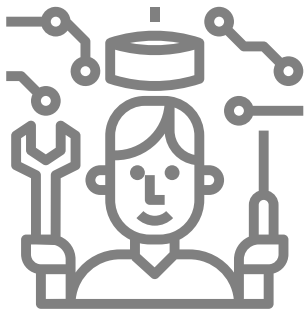
Name a reason why Python is better to use in data science instead of most other programming languages.

Naturally, Python is very rich in data science libraries, it's amazingly fast and easy to read or learn. Python's suite of specialized deep learning and other machine learning libraries includes popular tools like scikit-learn, Keras, and TensorFlow, which enable data scientists to develop sophisticated data models that plug directly into a production system.

To unearth insights from the data, you'll have to use Pandas, the data analysis library for Python. It can hold large amounts of data without any of the lag that comes from Excel. You can do numerical modeling analysis with Numpy. You can do scientific computing and calculation with SciPy. You can access a lot of powerful machine learning algorithms with the scikit-learn code library. With Python API and the IPython Notebook that comes with Anaconda, you will get powerful options to visualize your data.

Naturally, Python is very rich in data science libraries, it's amazingly fast and easy to read or learn. Python's suite of specialized deep learning and other machine learning libraries includes popular tools like scikit-learn, Keras, and TensorFlow, which enable data scientists to develop sophisticated data models that plug directly into a production system.

To unearth insights from the data, you'll have to use Pandas, the data analysis library for Python. It can hold large amounts of data without any of the lag that comes from Excel. You can do numerical modeling analysis with Numpy. You can do scientific computing and calculation with SciPy. You can access a lot of powerful machine learning algorithms with the scikit-learn code library. With Python API and the IPython Notebook that comes with Anaconda, you will get powerful options to visualize your data.



GENERAL TIPS

The most important things that you should remember for the beginning of your job interview are the definitions. If you have the definitions down and can explain them in an easily understandable manner, you're basically guaranteed to leave a good and lasting impression on your interviewers.

After that, make sure to revise all of the advanced topics. You don't necessarily need to go in-depth with each one of the thousands of data science interview questions out there. Revising the main topics and simply getting to know the concepts that you're still unfamiliar with should be your aim before the job interview.

Your main goal at the interview should be to show the knowledge that you possess. Whether it be interview questions for data analyst or anything else - if your employer sees that you're knowledgeable on the topic, he's much more likely to consider you as a potential employee.