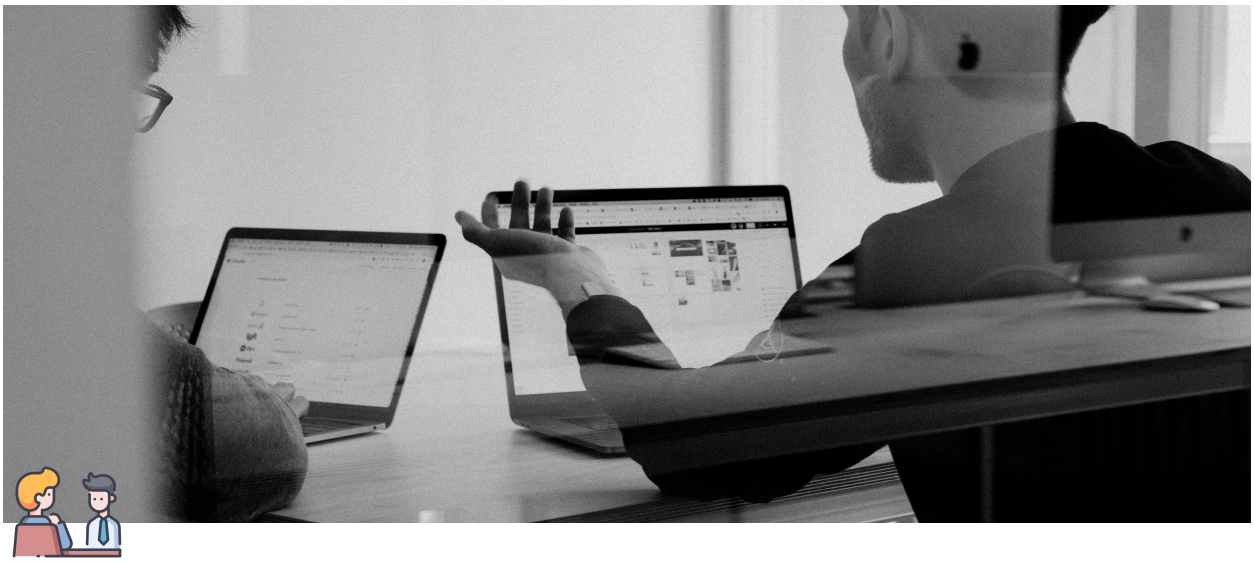

# Interview Questions

Tags

Questions:

## Generic

1. Which of your data analytic techniques are you proudest of?
2. How would you use handle outlier and unbalanced datasets?

## Basic ML

1. What are different feature-importance methods? when to use which?
2. Different between random-forest and LGB?
3. Difference between SVD and PCA?
4. Assumptions to do PCA?
5. Assumptions for Linear Regression?
6. Models we can use with very high-dimensional data?

7. Models we can use with very highly correlated data?
8. Difference between the back-propagation of neural networks and convolutional NN, and LSTM?
9. Different hypothesis testing? and when to use which? what's the p-value? what does a p-value of less than 0.05 mean?
10. What are loss functions for Linear and Logistics Regression and why are they so?

## Deep ML

1. Recommendation Question

```
Recommendation of products, based on the browsing history
of users

Data that we've

## ts.csv
this CSV contains the session name (first element) followe
d by a series of products browsed by user, and the last el
ement is the item purchased by user

s1 i1, i2, i3, -> i4 # Here in session s1 i1, i2 and i3 wa
s browsed by user and i4 was purchased
s2 i11, i22, i3 -> i45  # Here in session s2 i11, i22 and
i3 was browsed by user and i44 was purchased
...
We've one 1 crore rows in ts.csv


# item_features.csv
This csv contains encoded feature category and encoded fea
ture value for each item present in database.
```

```
i1 -> c1, v1; c2, v2 #item i1 has a feature c1 and corresp
onding feature value v1, likewise for c2 and v2
i2 -> c3, v5; c2, v11



There are total of 5000 items
There are max 100 categories c1, c2 ... c100
There are max 100 values v1, v2 .. v100

Given the browsing history of the user, recommend the prod
uct user will buy

i11, i12, i111, i1023 -> ??
```

2. Banned Keyword question

```
Suppose you're working for an e-commerce company where sel
lers from all over the country sell products on your syste
m, but some items can't be allowed to be sold on the platf
orm. For example, diamonds, weed, etc.

Suppose this country has over 40 local languages, and sell
ers can type a product in any of the languages.

Design an ML model that can classify given a phrase of ite
m name as 0/1, basically allowed/disallowed.

Challenges:
1. There can be spelling mistakes in the item name
2. There can be multiple languages in the same sentence, f
or example, "Blanche carpet", Blanche means white in frenc
h

A few examples, where 0 means not allowed, 1 means allow
```

```
# Ayurvedic medicine 0
# midicine 0
# shoes 1
# diamonds 0
# davai 0
# garlic paste 1
# paste 1
# pest 0


# The precision of the model must be very high, as the com
pany gets a severe penalty for allowing any of the dis-all
owed products;
```

## System Design

1. Design a stock-trading system, with the following requirements:
    a. Very high-frequency cache refresh from the server (What infra might you use?)
    b. Real-time stock purchase by clients (transaction commits need to happen immediately,  real-time)
    c. Personalize stocks based on the client’s past history and recommend similar stocks
2. Design a restaurant, with the following requirements:
    a. The restaurant should be able to serve at least a million concurrent users.
    b. Restaurants should be able to personalize recommendations real-time, i.e. if a user searches “Aloo Dum” similar dishes should be recommended, and this personalization should kick in within 10 seconds of the user searching for the dish.
        i. How will you capture interactions, what interactions will you capture and what features might you make out of it?

ii. how and where will you store it in real time?

iii. how will you use it to recommend people, and with what model?

c. The restaurant should refresh the prices of dishes every fixed interval (need you to design a system for this). This interval should be calculated for every dish specifically (as hitting API for refreshing dishes is costly, and prices of some dishes/beverages increase if ordered frequently like “Garlic Bread”, prices may increase/decrease every 15 minutes.)

## Spark

1. How to make a new column in a spark data frame?
2. how do you read data frames with more than 100 GB?
3. what is UDF? how to make a UDF with the spark-data frame? Ex, Add a number to a column

## SQL

```
# Supposed a table with the following field,

# table_a name
# person_id - varchar
# age - int
# gender - m/f
# exp_years - int


# table_b name
# person_id - varchar
# salary - int
# job_title - varchar
```

```
# Compute average salary based on job_title * exp
# 1. Group by job-title and show average-salary (for exp > 5)
# 2. For each-person, create value_index column (salary * exp)
```