Tableau:

Phone:

- snowflake schema vs star schema
- Python
    - decorators?
    - difference between list, sets and tuple
- Coding:
    - remove dups from list
    - remove symbols !,$,? from the string … use regular expressions
    - Reverse each line in file
    - SQL - not hard, but very confusing

Microsoft:

Phone:
Input 2 lists. For each element in list1 find closest element in list2. Maximum acceptable difference is 1 symbol. Example abc will match abd and acb but not cba.
[abc], [abc, abd] = [abc->abc]
[abc], [ffff] = []
[abc], [abd, abe] = [abc->abd] can pick any
[abc, bdr, aws, kk] , [abd, fjk, abec, ws, ka, kk] = [abc->abd, bdr->null, aws->ws, kk->kk]

GrubHub:

Phone:
Similar to:
https://www.geeksforgeeks.org/rat-in-a-maze-problem-when-movement-in-all-possible-directions-is-allowed/

- we have board with walls and we need to find a path from point A to point B
- implement point class, board class and fill board

Pinterest

Phone:

- Experience
- DWH design
- Hive performance tuning

On-site:

- Experience with drill down - what I did, why, what I would do differently
- Simple SQL, + coding:
    - https://www.baeldung.com/java-algorithm-number-pairs-sum
    - https://leetcode.com/problems/top-k-frequent-elements/
- Design DWH for pinterest metrics
- Design Facebook feeds (users, network, interests, post, activity...)

Dropbox:

Phone:

- SQL: Which employee has the highest salary per department?
- SQL: top 3 employees with highest salary per department
- https://leetcode.com/problems/group-anagrams/

Roku

Phone:

- SQL: simple, with windowing functions(example: print second largest salary in each department)
- find number in sorted array
- https://leetcode.com/problems/number-of-segments-in-a-string/

Spotify

Phone:

- How to handle skew
- How to delete element from the linked list, what complexity, how it is different from array
- Explain to child: map reduce, hash function, greedy algorithm
- What tasks are not good for map-reduce
- https://leetcode.com/problems/generate-parentheses/

Sqoop

Phone:

- Given a binary array, find the maximum number of consecutive 1s in this array.
  https://leetcode.com/problems/max-consecutive-ones/
  ```
  Input: [1,1,0,1,1,1]
  Output: 3
  Explanation: The first two digits or the last three digits are
  consecutive 1s.
      The maximum number of consecutive 1s is 3.
  ```

- Given a binary array, find the maximum number of consecutive 1s in this array if you can flip at most one 0
  ```
  Example 1:
        Input:  nums = [1,0,1,1,0]
        Output:  4

        Explanation:
        Flip the first zero will get the the maximum number of
  consecutive 1s.
        After flipping, the maximum number of consecutive 1s is 4.
  ```

USAA:

Phone:

- Experience questions
- DWH questions:
    - Star schema, Snowflake
    - OLTP vs OLAP
    - What is fact table
    - What is Dim table
- How Map reduce works
- DB SQL optimization, how to solve performance issues

Onsite:

1. Group exercise: you + 1 other candidate doing estimation 15min … after planning for r1 and r2 10min
2. Experience, 10 min for simple 10 sql(on laptop)
3. Design DWH for AGG table …. from source we getting transactions and clients. For reporting we need to prepare agg table with date, client attributes and amounts, trl count …. need to design from source to agg, what table we should create, write how to handle SCD2 and late arriving data
4. Python + CL
    a. input: ‘abcabcabc’, 3 -> output ccc …. print elements with step n
    b. *https://leetcode.com/problems/self-dividing-numbers/*

Lucid Holdings:

Phone(3h):

- Experience questions
- Deep dive into recent projects
- Design question: design pipline(get data, processing, storage, sharing) for word - count stream which should be aggregated and accessed via API(return each word count per day)
- DWH questions
- Many-to-Many Movies-Actors
- How to store hierarchy data
- Spark tuning questions

Onsite:

- Questions about experience
- Questions about streaming in AWS
- Design stream application with input: 1 minute batches with: survey - user - status output should be survey - status - number of users

Staples:

Phone:

- Experience questions
- Deep dive into recent projects
- TEZ vs MR
- Spark memory questions, questions about Catalyst
- How to calculate settings for Spark applications(memory, cores ….)

ThreadUP

Phone:

- Experience questions
- Deep dive into recent projects
- What issues I had with Hadoop and Spark
- Python: decorators, data types, generators
- ORC vs Parque
- CAP theorem

NerdWallet:

Phone:

- Experience questions
- Simple SQL + windowing functions
- Coding - print nested list: word - index(+index from nested list) - value
  Input: List = [1,2,[3,4]], word: Foo
  Output:
  Foo - 0 - 1
  Foo - 1 - 2
  Foo - 2 - 0 - 3
  Foo - 2 - 1 - 4

Lyft

Phone:

- написать компаратор версий а потом используя его написать сортировку (можно самую простую аля селекшн/инсершн/бабл)
- Write a method to check if input string has balanced parentheses
  Input 1: ( )[ ] { }( )   output: balanced
  Input 2 : ([{] }) output : not-balanced

- Сумма двух двоичных:
  Input: a = "11", b = "1"
  Output: "100"

  Input: a = "1010", b = "1011"
  Output: "10101"
- SQL

  ```
  order_tbl:
  (
  customer_id bigint,
  product_id bigint,
  order_id bigint,
  price decimal,
  purchase_dt date
  )

  product_tbl:
  (
  product_id bigint,
  product_name varchar,
  desc text,
  default_price decimal
  )
  ```

  Note: price in order_tbl can be null, in that case price paid for product is product_tbl.default_price

  - Q1: Get number of orders for each product per day
  - Q2: Get unique list of customers that bought products over any two consecutive days

Onsite:
Моделирование:
 - надо было придумать KPI чтоб оценивать єффективность продукта
 - потом нарисовать какие данньіе, на каком єтапе и в каком формате тьі хранил бьі (raw - stg - dm - reporting layers)

DB fundamentals:
 - поговорили про опьіт
 - поговорили о продукте и о том какие данньіе я б логировал
 - задачки на SQL, не сложньіе (analytical functions)

CS fundamentals:

- несложная задача (dictionary + list в python), но собеседователь говнюк немного попался, постоянно перебивал и сбивал

Design:

- Design whatsapp like service
- надо было задизайнить систему в которой от источника до дешборда стриминг данных обрабатьвался б и приходил с задержкой не больше минуты + заимплементировать кусок в котором данные обрабатываются

(я его провалил - не знаю ни классов ни методов для обработки стриминга ни в Spark ни в Beam)
не думаю, дизайн я завалил полностью

Nortal

Tech:
Diff between SQL and NoSQL (schema free)
Diff between LEFT and RIGHT join
Diff between transactional DB and DWH
How to optimize in transactional DBs
Main challenges for data engineer
How to scale relational DB (SQL Server)
Vertical and Horizontal partitioning

Zillow

Phone:
Non-tech:
What last product you worked on?
What different teams you coordinated with (not only eng)?
Tell me about situation when you work with a person you didn't like;
Tell me about a product you like; What can they do better? Why are they better than their competitors?
Tell me about a situation when customer asked to do something but you thought it should be done differently;
What was the last thing you've learned in Product management;

Tech:
Difference between INNER and FULL join;

Onsite:
Tech:
Very simply SQL questions: GROUP BY, JOINS
Give example of a pipeline you've build

How would you create technical specification for a system to manager elevators in luxury apartment complex

Non-tech:
How you handled interaction with difficult people
What if in the meeting with other PMs someone said that your product has an issue
How would you motivate your team members
Are you ok that the product you'll be supporting is for internal use and not for end users of Zillow

Amazon

From Eugene
Phone:

- Experience
- Simple SQL, no windowing functions
- Python: https://leetcode.com/problems/fizz-buzz/
- How would I store orders/canceled/modified orders
- What metrics I would track and analyze for the Alexa

On-site:

- SQL - all with windowing functions (lead, cumulative sum, rank)
- Python - open file, remove columns
- SCD-2 in hive
- SQL Performance tuning
- Designa amazon music - what should I do to increase the number of created playlists
- Design DWH for online games(Casinos, Bingo, bets ...) - high level
- How do you handle tasks with conflicting priorities
- Tell me about largest projects you worked on - what you will do differently
- Have you ever had a conflict with your manager? How did you resolve it?
- How did you handle negative feedback - give example
- What is not in your resume
- how you learning new things
- Example where I was wrong
- Provide example when your teammate refused to help - how did you handle it

-----
Non-tech:
What was the last thing you've learned

Tech:
(Hadoop) What are the components of Hive (MR, MetaStore);
(Hadoop) File formats - ORC, Parquet
(Hadoop) How do you sync data between platforms - Sqoop, BTEQ

(Spark) Can you choose Spark as execution engine in Hive;
(Spark) What is the difference between RDD and DataFrame;

Netflix

Tech:
If new data is coming how would you create a pipeline;
When to use star vs snowflake;
What technology you learned last time;
How would you rate from 1 to 5 knowledge in sql, spark, python;

Non-tech:
When you had a disagreement with your boss, what did you do;
When you had a disagreement with business, what did you do;
Why you decided to be a data engineer;
What was last challenging work;

Whitepages

Tech:
Write a code to compare two strings if they have the same set of characters;
You have a personal information and want to create a shortest possible link for each person. Each person has a first name, last name, country, state, city. In case person is unique for all countries, link should be only by first name and last name. If he is unique on a country level then first+last+country, etc.

Onsite:
1. Find the biggest subset of characters in 2 strings:
abcdefg
aabcdgf
bcd is the biggest one

2. Want to store 100 records in cache and replace them with new data from request if not exists:
Bound = 100
Get(k): V/null
Set(k,v): -

Hulu

From Eugene:

- What metrics I would have for search on Hulu site, how I would stor data to support analytics for search process on hulu
- Recalculating/restating daily snapshots in Spark
- Implement Join(any) in python: input 2 arrays: output result of inner join.

phone screening (M)
Tech:
1. (Modeling) Design a database to support a Spotify-like media cataloging application and activity. The application should catalog artists, their albums, songs in each album, playlists, users that created/shared various playlists. Your database must store data described by every entity, relationship. Your answer should consist of the table and column names. Assume the following about the application: Every song belongs to some album, every album was released by an artist, and every playlist was created by some user.

2. (SQL) What are the top 10 songs which have been added to the most playlists?
3. (SQL) List all the artists which do not have any songs in users' playlists.
4. (SQL) Find the top 5 playlist per user (for instance 0 playlists - 10,000 users; 5 playlists - 8,700 users; 2 playlists - 5,600 users; 12 playlists - 1,200 users)

Onsite:
How to convert to SCD type 2
Find unique users who watched tv show; find same but for a week rolling every day;
Call center to sell credit cards: find top 10 sellers for top 10 banks
Your most interesting project
Most proudest moment (and go into tech details why, what you've implemented)
What is important in team culture
What do you like/dislike in scrum

Phone: (R)
How to Select data without duplicates
Diff between DWH and datamart
What is a granularity of a table
What are factless fact tables? Give an example, what is the use case?
What is cross join, when to use?
What is junk dimension? (you can build it using cross join)
How would you consume unstructured data?
How would you set up SLA?
Would you join 2 fact tables?

What will you work on in data if you have free time (things that always there but no one have time to work on them)?
Some data (subscriptions) is coming on an hourly basis, other (customer's data) is coming on daily basis. How would you consume it, what you'll tell business?
Need to build fact table which will show unique number of users watched something new on a weekly basis. Data is coming on a daily basis though.
How would you implement SCD Type 2 in Hadoop?

Onsite:
Number of distinct users by week;
Same, but cumulative;
How to backfill that data;
What performance issues you can come up with;
Rewrite sql with window function but not using window function;

General

Amazon principals - must have
Five years bulshit
Tell me about your self
Challenging you manager
Why do you want to work in …
when to escalate
How do you handle tasks with conflicting priorities
Tell me about largest projects you worked on
The least exciting thing in your job
Example of the decision you made with lack of information
How do I manage risks
How do you encourage customers to use a new product(Amazon)
What health metrics you are checking on daily basis … and what metrics you would use for messaging system
How do you work with difficult people?
Describe some project management approaches that you use(strengths and weaknesses of each)
How did you manage missed deadlines
What tools did you use for your PM work
How would you improve your product?
How you developing your technical skills
How do you track delivering quality in your product?