Data Contractor Interview Test

The test has two components. This is the SQL-focused portion, testing your fundamental skills and ability to work with other people's code. Working in Riot's data systems, you will be expected to have strong SQL fundamentals. Query syntax, functions, structure, etc. Each question has specific answer criteria, but there may be multiple correct approaches. Save your answers in some text format.
You can leverage http://sqlfiddle.com/ if you want to build the schemas and data to test your queries.

The second component of the test focuses on Excel and pivot tables. In a separate excel sheet, you will be given some sample data and a series of data questions. More instructions are contained in that sheet.

Feel free to consult Google or whatever other resources you would have access to in a normal workday. No 2-way communication about the problems should be allowed. Save your answers and send them over within 24 hours. I will review your answers and talk through them in a live phone call afterwards, so you can convey more thought or nuance then.

SQL Fundamentals Test

1) Library Books

Write a query to list the name of every library in the state of KS and the number of books it has.

Book

ISBN	Author	Library_id
5634	Bill Nye	1
5103	Elon Musk	1
4294	Bill Clinton	2
5435	Bill Gates	2
7773	Bob Sagat	2
6456	Ned Sanders	3
5879	Homer Simpson	3
5645	Xin Luo	3
7774	Donald Trump	3

Library

Library_id	Name	State	ZipCode
1	KU College	KS	78543
2	Hollywood Public	CA	90210
3	KS Library	KS	78501
4	Bobs Library	KS	78566

2) Second Transaction

An analyst writes the following query, attempting to find the amount each account spends on their second purchase. The query returns unusual results, so they come to you for review:

```
SELECT
account_id
, MIN(purchase_date) AS second_purchase
, MAX(amount)
FROM transactions t
WHERE purchase_date NOT IN (
SELECT MIN(purchase_date)
FROM transactions m
)
GROUP BY account_id;
```

What does this query achieve? What are some limitations of this query? How would you improve this query? Please write your improved query.

3) Account Login Data

I am trying to find the account that has logged in for the longest aggregate time for each region. I wrote the following query but it's not working properly. Explain to me why the query fails, and write a query that works and that only return the desired outcome.

```
SELECT realm_id, account_id, total_loggedin_seconds
FROM platform_accounts
WHERE total_loggedin_seconds = max(total_loggedin_seconds)
GROUP BY realm_id, account_id;
```

Platform Accounts table
Summary stats for accounts. One row per each game session.

platform_accounts

realm_id	account_id	total_loggedin_seconds
1	1234	1520
1	412	523426
1	623	10
1	1234	62345
2	412	1623

4) Survey Data

I am creating a table to capture some player data,
which involves a short 3-5 question true/false survey on their player experience.
Currently I'm using the following design, but I am concerned because my coworker has been bringing up the idea of expanding the number of questions in the future, which would break this schema. What can I change to accommodate extra questions? What are the pros/cons of your solution?

realm_id	account_id	survey_date	q_1	q_2	q_3	q_4	q_5
1	153	2015-12-30	T	T	F	T	F
2	122	2015-01-05	F	F	F	NULL	NULL
2	152	2015-03-10	T	T	F	T	NULL

5) Jinx and Vi

Q: Write a query to get the account that wins the most Jinx games each day.
Q: Write a query to get a list of all games where Vi and Jinx are both played on opposite teams.

Player Game table.
One row per player game

player_games

account_id	game_id	game_date	game_type	champion	team_id	win_flag
165234	11543276	2016-01-01	SR_NORMAL	Garen	1	0
175245	12365164	2016-01-02	SR_RANKED	Vi	2	0
613472	27345231	2016-01-03	SR_NORMAL	Vi	1	1
652153	27345231	2016-01-03	SR_NORMAL	Jinx	2	0

6) Player Metrics by Champion

Now suppose we have another table that summarizes how much each player plays and spends each month. We are interested to see if Jinx players tend to spend more or less than players of other champions. Your coworker writes the following query, but it doesn't seem to work.

Rewrite the query to get the average games played and usd spent for each champion.
What was wrong with the original query?

player_monthly_summary

account_id	month_date	games_played	usd_spent
165234	2016-01-01	3	NULL
165234	2016-02-01	14	10
613472	2016-02-01	42	NULL
652153	2016-02-01	13	5

```
SELECT
      champion
      , month_date
      , count(distinct account_id)   as accounts
      , sum(games_played)      as games_played_total
      , sum(usd_spent)   as usd_spent_total
      , sum(games_played)/count(distinct account_id)    as
games_played_per_account
      , sum(usd_spent)/count(distinct account_id)       as usd_spent_per_account
FROM  player_monthly_summary   pms
JOIN  player_games       pg
ON    pg.account_id = pms.account_id

GROUP BY champion
```

7) Debug Query

The following query has a lot of errors and will not run. Find as many as you can.

```
-- among players who play neither Jinx nor Vi for that month, how
much do they play or spend?
SELECT
     account_id
     , month_date
     , games_played
     , CASE WHEN usd_spent = NULL THEN 0 ELSE usd_spent
FROM player_monthly_summary pms
JOIN (
     -- monthly Jinx players
     SELECT
          account_id
          , date_trunc(month, (game_date) as month_date
          , champion
          , count(*) as jinx_game_count
     FROM player_games
     WHERE champion = "Jinx"
     GROUP BY account_id, game_date
) pg
ON   pms.account_id = pg.account_id
     AND pms.month_date = pg.game_date

JOIN (
     -- monthly Vi players
     SELECT
          account_id
          , date_trunc(month, (game_date) as month_date
          , champion
          , count(*) as vi_game_count
     FROM player_games
     WHERE champion = "Vi"
     GROUP BY account_id, game_date
) pg
ON   pms.account_id = pg.account_id

WHERE jinx_game_count = 0 or vi_game_count = 0
```

8) Account Sampling

An analyst is trying to take a 1000-person sample from the accounts table in Vertica. They write the following query, and it gives them the first 1000 accounts created. This will not do, since it is not a representative sample.

```
SELECT
     *
FROM db.platform_accounts
LIMIT     1000
```

The analyst iterates on the query to randomize the sample. They write the following query. What is the query doing? Will it work? Is there a better way to do this? If yes, write your query.

```
SELECT
     *
FROM (
     SELECT
          *
          , random() as rand_sample
     FROM db.platform_accounts
) pa
WHERE     rand_sample < 0.1
LIMIT     1000
```

9) String Detection

How do you identify the comments that contain the word “noob” and its variation? For example, we’d like to capture the comments with “n00b”, “noobs”, and “NOOOOB”.

How would you identify phone numbers contained in a large text file? Suppose that phone numbers may have varying formats and may or may not include area codes (e.g. 1-234-567-8900, (234) 567 8900, 567.8900, etc.).

10) Magnitude Estimation

The following questions are broad estimations. We do not expect a precise answer, but a general estimate. Write up your reasoning behind your estimate.

How many people in your hometown own cars?

How many Cheerios are in a box?

For LoL players only:
How many Teemo shrooms are placed each day?

For Non-LoL players only:
How many Pizzas are delivered in Los Angeles each week?

.