Test exercise for the backend developer

Imagine you have a database with two tables A ("pageviews") and B ("atc_clicks"), each table has about 1M of rows in it and any row in A may have zero or many related rows in B.

Your task is to fetch all rows from B joined with rows in A, process them on the backend and calculate aggregated result.

Table A (aka "pageviews")

Stores info about every visit on product description page)

Columns: id, time, product_id, visitor_id, browser_name, url.

Table B (aka "atc_clicks")

Stores raw data about clicks on "Add to cart" button). Linked to table A via *impression_id*.

Columns: id, impression_id, click_id, local_time.

Tasks

Populate

Please write a script that will populate tables A and B with nearly random data, assuming you should get 1 million of rows in every table. There should be around 700'000 unique visitor_id and about 200'000 unique product_id in table A.

For **url** column please use hostnames from this set: ["localhost", "127.0.0.2", "google.com", "shop1.com", "shop2.com", "www.shop1.com", "www.google.com", "shop4.ru", "www3.shop4.ru"].

And append any valid random path to it so that you get massive amount of valid random URLs that are from specific hostnames. Schema (https / http) should also vary randomly.

About 50% of rows in the table A should not have related rows in table B. Every row in B must have only one row in A (one click can't happen on several page views).

Some impressions (table A) will have between 1 and 6 clicks in table B - decide that randomly.

Define time for every row in A randomly between 1.1.2017 and 31.08.2017. Time for B should be shifted from A.time to any reasonable random interval into the future (clicks happen after page view).

SQL

Please write SQL queries that prove that you populated the database according to the requirements.

How many unique products were visited in June? Please write SQL query for that.

Share the queries via GitHub.

Fetch data

Now you have almost real data about product page visits and clicks. Let's calculate something useful.

Please write JS script that calculates the following:

Let's define a "conversion" as a unique combination of product_id and date for the given shop (one shop means hostnames like www.shop.com a nd shop.com).

How many conversions are there for shop1.com (including www. version) that were made in July?

Here you should fetch data from the DB, filter out something and make calculations on the backend.

Feel free to modify data structure as you think necessary.