Задание по python+pandas+SQL

Вам будет предоставлен доступ к базе данных с двумя таблицами: обращения и сессии. Нужно написать скрипт на языках python и sql, который для каждого обращения двумя способами находит источник – рекламную кампанию, которая к нему привела.

Результатом работы будут являться два файла:

* Файл самого скрипта .py.
* Файл requirements.txt со списком пакетов и их версий, использованных при работе (для воспроизводимости результата).

Подробности по выполнению задания:

* Вам на почту или другим способом была выслана пара логин-пароль.
* Сервер postgresql расположен по адресу. Подключиться нужно к базе данных data.
* В БД лежат две таблицы:
* таблица web\_data.sessions, в которой хранится информация о сессиях 1 000 случайно отобранных посетителей двух сайтов, произошедших за небольшой период в 2018 году;
* таблица web\_data.communications, в которой хранится информация обо всех обращениях этих пользователей, произошедших за тот же период.
* Напишите скрипт на языке python, который решает задачу двумя способами:
* Формирует SQL-запрос, который решает задачу, отправляет его на сервер и скачивает результат выполнения в pd.DataFrame (в данном пункте оцениваются знания SQL);
* Получает обе таблицы из БД целиком (SELECT \* FROM web\_data.communications; SELECT \* FROM web\_data.sessions;) в два объекта pd.DataFrame, затем решает задачу в pandas (в данном пункте оцениваются знания pandas).
* Задача, которую нужно решить двумя способами (с помощью SQL и pandas). Нужно получить список всех обращений. Для каждого обращения (communication\_id) найти одну – последнюю на момент обращения – предшествующую сессию данного посетителя на том же самом сайте, что и обращение. Кроме того, нужно указать, которая это по счету сессия данного посетителя на данном сайте. Наконец, если для какого-то обращения нет ни одной предшествующей сессии, его (обращение) все равно нужно вернуть с пустой информацией о сессии.

Например, если сессий у пользователя на одном сайте было три: 3, 4 и 6 апреля 2018 года, а обращение одно – 5 апреля, то нужно найти сессию от 4 апреля и указать, что она была второй по порядку.

В конечной таблице должны быть следующие поля: communication\_id, site\_id, visitor\_id, communication\_date\_time, visitor\_session\_id, session\_date\_time, campaign\_id, row\_n (порядковый номер сессии этого пользователя в рамках этого сайта).

* Ожидается, что результаты решения обоими способами будут совпадать друг с другом - в качестве необязательного задания можете сравнить в pandas результаты, полученные двумя способами.