Мы попробуем использовать возможности Spark для анализа данных clickstream пользователей новостного Интернет-портала.

a.       Создайте схему будущего фрейма данных. Схема должна включать следующие атрибуты:

   id -  уникальный идентификатор посетителя сайта. Тип – последовательность чисел фиксированной длины. Данное поле не является первичным ключом.

   timestamp – дата и время события в формате unix timestamp.

   type – тип события, значение из списка (факт посещения(visit), клик по визуальному элементу страницы(click), скролл(scroll), перед на другую страницу(move)).

   page\_id – id текущей страницы. Тип - последовательность чисел фиксированной длины.

   tag – каждая страница с новостью размечается редакцией специальными тегами, которые отражают тематику конкретной новости со страницы. Возможный список тематик: политика, спорт, медицина и т.д.

   sign – наличие у пользователя личного кабинета. Значения – True/False.

b.       Создайте датафрейм с описанной выше схемой данных.

c.       Наполните датафрейм данными. Пример:

(12345, 1667627426, "click", 101, "Sport”, False)

d.       Решите следующие задачи:

   Вывести топ-5 самых активных посетителей сайта

   Посчитать процент посетителей, у которых есть ЛК

   Вывести топ-5 страниц сайта по показателю общего кол-ва кликов на данной странице

   Добавьте столбец к фрейму данных со значением временного диапазона в рамках суток с размером окна – 4 часа(0-4, 4-8, 8-12 и т.д.)

   Выведите временной промежуток на основе предыдущего задания, в течение которого было больше всего активностей на сайте.

   Создайте второй фрейм данных, который будет содержать информацию о ЛК посетителя сайта со следующим списком атрибутов

1.       Id – уникальный идентификатор личного кабинета

2.       User\_id – уникальный идентификатор посетителя

3.       ФИО посетителя

4.    Дату рождения посетителя

5.       Дата создания ЛК

   Вывести фамилии посетителей, которые читали хотя бы одну новость про спорт.

   **Выведите 10% ЛК, у которых максимальная разница между датой создания ЛК и датой последнего посещения.**

   **Вывести топ-5 страниц, которые чаще всего посещают мужчины и топ-5 страниц, которые посещают чаще женщины.**

e.       **Создайте в Postgres таблицы аналогичной структуры и выполните следующие задания с помощью Spark.**

   **Создайте витрину данных в Postgres со следующим содержанием**

1.       **Id посетителя**

2.       **Возраст посетителя**

3.       **Пол посетителя (постарайтесь описать логику вычисления пола в отдельной пользовательской функции)**

4.       **Любимая тематика новостей**

5.       **Любимый временной диапазон посещений**

6.       **Id личного кабинета**

7.       **Разница в днях между созданием ЛК и датой последнего посещения. (-1 если ЛК нет)**

8.       **Общее кол-во посещений сайта**

9.       **Средняя длина сессии(сессией считаем временной промежуток, который охватывает последовательность событий, которые происходили подряд с разницей не более 5 минут).**

10.   **Среднее кол-во активностей в рамках одной сессии**

f.        **Редакция совместно с аналитиками хотят провести масштабную рекламную кампанию, в рамках которой на сайте будут выпущены 3 новости с новой тематикой. Рекламный бюджет позволяет охватить только 10% доступной вам аудитории посетителей сайта. Напишите запрос, который позволит вычислить целевую аудиторию для данных новостей. Постарайтесь объяснить ваше решение.**