ПРОЕКТ 3

АНАЛИЗ ЛОГОВ

ЮЛИЯ ХАБИНА

ОПИСАНИЕ ПРОЕКТА

- Журналы логов содержат очень подробную информацию о действиях пользователей.
- Просмотр страницы может содержать много строк журнала, а сеанс может состоять из нескольких просмотров страниц.
- Другой важной характеристикой лог-файлов является то, что они обычно являются очень большими, и это нужно учитывать.
- Поэтому анализ логов может дать ценную информацию бизнесу о действиях пользователей.

ЦЕЛИ ПРОЕКТА

- Разработать скрипт формирования витрины следующего содержания:

1. Суррогатный ключ устройства
2. Название устройства
3. Количество пользователей
4. Доля пользователей данного устройства от общего числа пользователей
5. Количество совершенных действий для данного устройства
6. Доля совершенных действий с данного устройства относительно других устройств
7. Список из 5 самых популярных браузеров, используемых на данном устройстве различными пользователями, с указанием доли использования для данного браузера относительно остальных браузеров
8. Количество ответов сервера, отличных от 200 на данном устройстве
9. Для каждого из ответов сервера, отличных от 200, сформировать поле, в котором будет содержаться количество ответов данного типа

ПЛАН РЕАЛИЗАЦИИ

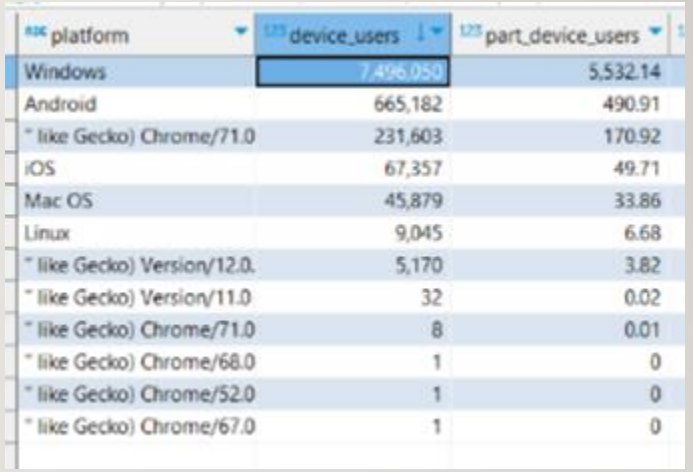
- 1 этап – распарсить лог: вначале необходимо подобрать регулярные выражения на тестовом образце, потом провести на всем датафрейме.
- 2 этап – вычислить необходимые данные
- 3 этап - записать витрину в таблицу.

ИСПОЛЬЗУЕМЫЕ ТЕХНОЛОГИИ

- Так как журнал логов огромный, содержит более 10 млн записей, то для обработки данных будет использоваться Spark
- Анализ показателей будет проводиться в Jupiter Notebook при помощи стандартных библиотек (pandas, numpy)
- Для записи данных будем использовать psycopg2
- Для хранения витрины создадим таблицу в PostgreSQL.

РЕЗУЛЬТАТЫ РАЗРАБОТКИ

- Тетрадь с кодом
- Файл с результатами Log_datamart.csv
- Таблица Log_mart в базе данных Log



platform	device_users	part_device_users
Windows	7,496,050	5,532.14
Android	665,182	490.91
like Gecko) Chrome/71.0	231,603	170.92
iOS	67,357	49.71
Mac OS	45,879	33.86
Linux	9,045	6.68
like Gecko) Version/12.0.	5,170	3.82
like Gecko) Version/11.0	32	0.02
like Gecko) Chrome/71.0	8	0.01
like Gecko) Chrome/68.0	1	0
like Gecko) Chrome/52.0	1	0
like Gecko) Chrome/67.0	1	0

ВЫВОДЫ

- Прежде чем, приступать к работе с файлами лучше всего изучить немного предметную область и понять, что обозначает каждый столбец/запись. Казалось бы это очевидно, но так просто об этом забыть;
- При работе с большими файлами стоит оценить мощности компьютера и обкатать алгоритм вначале на файле с небольшим количеством строк;
- Для быстрых расчетов лучше не злоупотреблять командой show(), иначе можно очень много времени потерять на этом;
- Промежуточные файлы с расчетами лучше сохранять в файл, чтобы в случае ошибок не терять время на повторные вычисления.
- Промежуточные переменные лучше писать в отдельных ячейках, а в итоговом расчете комментировать, делая только те вычисления, которые нужны для витрины для экономии ресурсов.