# ΠΡΟΕΚΤ 3 ΑΗΑΛИ3 ΛΟΓΟΒ

Климантов Артём (Lapis)

#### ЦЕЛИ ПРОЕКТА

- 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

#### ОСНОВНЫЕ ПРОБЛЕМЫ ПРОЕКТА

- Каждая запись в файле хоть и содержит не так много информации, но из-за большого количества записей, а именно 10365152 различных строк, анализ файла как минимум не возможен на слабых кластерах, а как максимум занимает около 40 минут, что не является хорошим показателем если хочется проверять гипотезы.
- Для приемлемых сроков создания и отладки основного скрипта для парсинга пришлось ((отрезать)) небольшой кусок в 5к строк и проверять работу на нём, так как мой компьютер хоть и не является слабым, но в силу некоторых железных ограничений выполнить преобразования на полном файле не представляется возможным

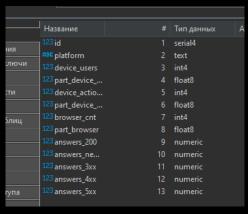
#### ПЛАН РЕАЛИЗАЦИИ ПРОЕКТА

- Распарсить текстовые строки на нужные данные с использованием регулярных выражений и парсинга через методы библиотеки httpagentparser
- 2. Вычислить необходимые количественные метрики
- 3. Создать и заполнить витрину данных в соответствии с запросами задачи

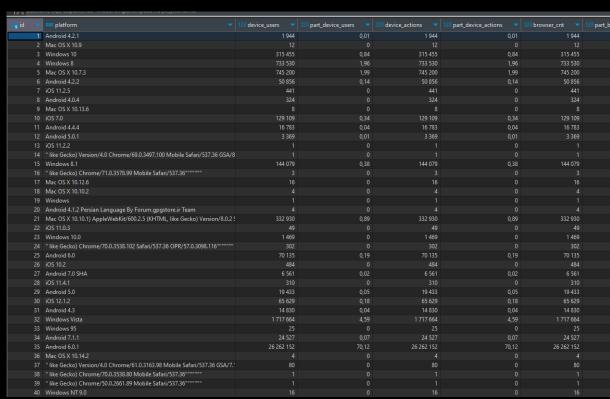
### СТЕК ТЕХНОЛОГИЙ

- В силу объемности изначального журнала данных в виде обработчика был выбран Spark и его имплементация для питона PySpark.
- Анализ и расчёт всех необходимых метрик проводился с использованием Jupyter и библиотек Pandas и PySpark.
- Парсинг строк производился с использованием регулярных выражений и библиотеки httpagentparser
- Для передачи информации в базу данных использовался psycopg2.
- Для хранения витрины была создана таблица в PostgresSQL.

- Jupyter Notebook содержащий необходимые ячейки после запуска которых в базе данных будет создан и заполнена таблица с метриками.
- Файл с данными витрины df\_mart.csv
- Витрина данных Log\_mart



#### РЕЗУЛЬТАТ



## ВЫВОДЫ В ХОДЕ ВЫПОЛНЕНИЯ

- 1) Самое важное это верно оценить мощность своего компьютера как с точки зрения таких комплектующих как оперативная память и количество ядер процессора, так и с точки зрения свободного места для создания кэша и прочего.
- 2) Jupyter очень хитрый и если переменная нигде не используется то до момента востребованности он ее даже трогать не станет, данная особенность сыграла со мной очень злую шутку, которая закончилась очень резким и неожиданным синим экраном смерти винды.
- 3) Show() очень странный метод, который может «выплюнуть» 100 строк за секунду, а может над тремя думать пять минут
- 4) Если есть возможность, то обязательно сохранять промежуточные данные, ибо считать 20 минут каждый раз очень больно и долго.
- 5) Чем меньше лишнего тем лучше, не стоит плодить ворох переменных на любую свою идею, все лучше тестировать в изолированной среде.