

# ИТОГОВЫЙ ПРОЕКТ по программе «Инженер данных»

## Проект № 3. Анализ логов

Выполнила:  
Лазарева  
Ирина Михайловна

2022

# Название проекта: Анализ логов

## Описание проекта:

Разработать скрипт формирования витрины следующего содержания:

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

Источник данных: <https://disk.yandex.ru/d/BsdiH3DMTHpPrw>

# Цели проекта, бизнес-задачи, требования

- Цель: создать скрипт для формирования витрины на основе логов web-сайта.
- Необходимо проанализировать лог-файл с информацией о посещении сайта пользователя и ботами, которая позволит составить более точную и подробную статистику для того, чтобы понять, откуда приходят пользователи, где они находятся и какими устройствами пользуются для визита.
- Благодаря данному лог-файлу нужно получить информацию об используемом устройстве, браузере, IP-адрес посетителя и его действиях.

# План реализации

1. Постановка задачи.
2. Анализ данных в предоставленных файлах.
3. Проектирование схемы данных для формирования требуемой витрины.
4. Анализ подстроки user-agent и формирование файла, содержащего необходимую информацию.
5. Создание на основе лог-файла таблицы для размещения необработанных данных.
6. Процесс data quality для анализа данных на корректность, исправление ошибок/опечаток, определение структуры и типов данных.
7. Формирование базовой таблицы.
8. Разработка запросов для формирования требуемых таблиц в соответствии со схемой данных.

# Используемые технологии

- Система виртуализации VirtualBox с операционной системой Ubuntu. Достоинства: opensource и бесплатно.
- Система контейнеризации Docker, сборка Apache Spark Standalone Cluster on Docker (Spark Cluster, JupyterLab). Достоинства: легко масштабируется на решение реальных задач; opensource и бесплатно.
- Инструмент для написания и отладки кода JupyterLab (PySpark). Достоинства: привычный инструмент; opensource и бесплатно.

# Структура исходных данных

1. Данные файла access.log сохраняются в следующей схеме:

```
schema = T.StructType(fields=[
    T.StructField("IP", T.StringType(), True),
    T.StructField("sign_1", T.StringType(), True),
    T.StructField("sign_2", T.StringType(), True),
    T.StructField("Date_access", T.StringType(), True),
    T.StructField("Date_access_", T.StringType(), True),
    T.StructField("Action", T.StringType(), True),
    T.StructField("Status", T.IntegerType(), True),
    T.StructField("Size", T.IntegerType(), True),
    T.StructField("sign_3", T.StringType(), True),
    T.StructField("User_agent", T.StringType(), True),
    T.StructField("sign_4", T.StringType(), True),

    T.StructField("Browser", T.StringType(), True),
    T.StructField("Device", T.StringType(), True),
])
```

2. Структура базовой таблицы для формирования витрины:

IP	Date_access	Action	Status	Size	User_agent	Browser	Device
54.36.149.41	[22/Jan/2019:03:5...	GET /filter/27 13...	200	30577	Mozilla/5.0 (comp...	AhrefsBot	Spider
31.56.96.51	[22/Jan/2019:03:5...	GET /image/60844/...	200	5667	Mozilla/5.0 (Linu...	Chrome Mobile	Huawei
31.56.96.51	[22/Jan/2019:03:5...	GET /image/61474/...	200	5379	Mozilla/5.0 (Linu...	Chrome Mobile	Huawei

# Описание результирующих таблиц

Таблица 1: Устройства по пользователям (Devices\_Users)

Атрибуты	Имя атрибута	Тип значений
Суррогатный ключ устройства	id_device	целый
Название устройства	Device	строковый
Количество пользователей	Count_Users	целый
Доля пользователей данного устройства от общего числа пользователей	Ratio_Users	вещественный

Таблица 2: Устройства по действиям (Devices\_Actions)

Атрибуты	Имя атрибута	Тип значений
Суррогатный ключ устройства	id_device	целый
Название устройства	Device	строковый
Количество совершенных действий для данного устройства	Count_Actions	целый
Доля совершенных действий с данного устройства, относительно других устройств	Ratio_Actions	вещественный

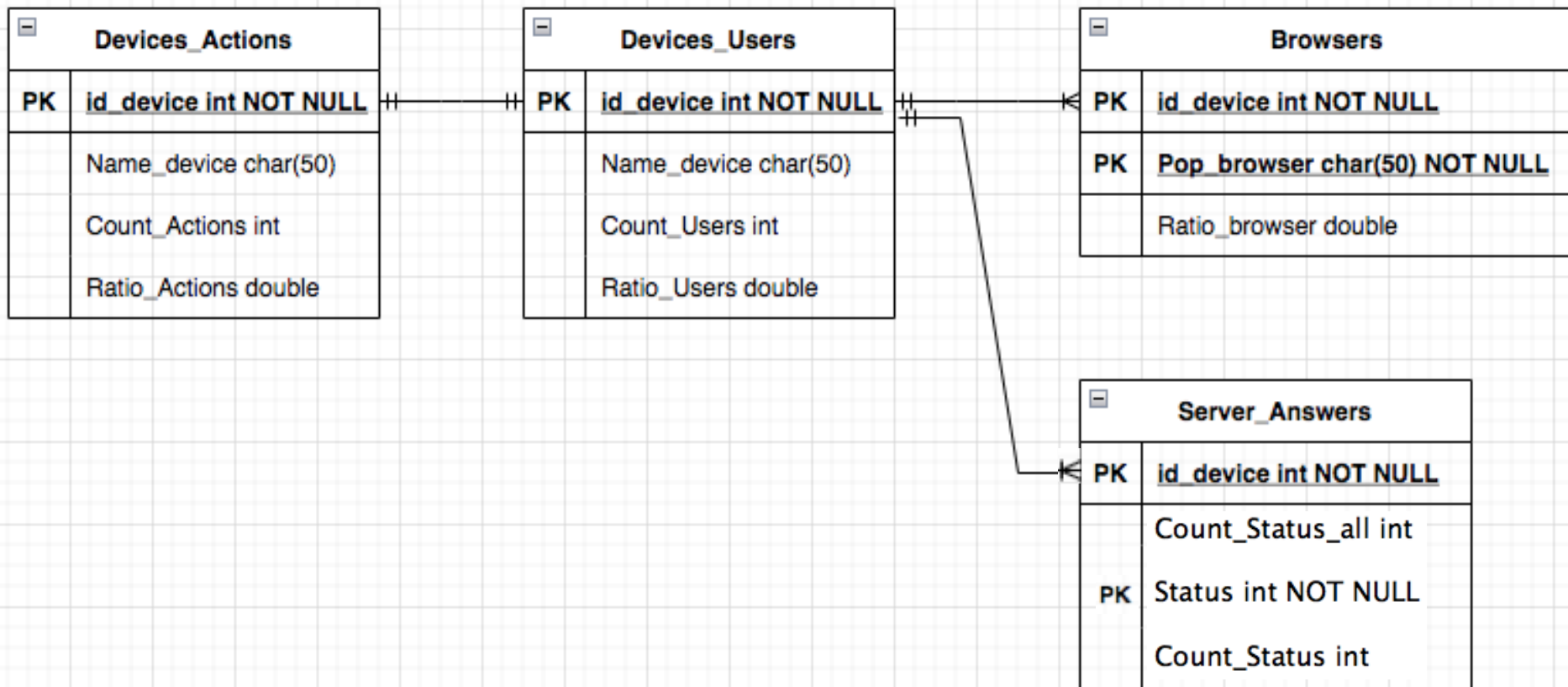
Таблица 3: Браузеры (Browsers)

Атрибуты	Имя атрибута	Тип значений
Название устройства	Device	строковый
Популярный браузер	Browser	строковый
Доля использования данного браузера относительно остальных браузеров	Ratio_browser	вещественный
Рейтинг браузера	row_number	целый

Таблица 4: Ответы сервера (Server\_Answers)

Атрибуты	Имя атрибута	Тип значений
Название устройства	Device	строковый
Общее количество ответов сервера отличных от 200	Count_Status_all	целый >
Статус ответа отличный от 200	Status	целый
Количество ответов сервера	Count_Status	целый

# Схема хранения данных в таблицах





# Результаты разработки

- Создана таблица (датафрейм) для хранения необработанных данных
- Импортированы полученные из сети данные
- Проведена проверка на корректность полученных данных
- Удалены неинформативные столбцы
- Проведен анализ значений полей строки лог-файла. Из подстроки «User\_agent» выделены значения «Device.Brand» и «Browser.Family» и добавлены в таблицу
- Сформирована базовая таблица (датафрейм) с результатами предобработки исходных данных
- На основе запросов к базовой таблице сформированы четыре спроектированные таблицы

# Выводы

- Поставленная задача выполнена в части проектирования, предобработки исходных данных и реализована в коде
- Возможности выбранных технологий достаточны для решения поставленной задачи
- Полученные витрины данных сохранены в трех файлах с целью уменьшения дублирования