



Курсовой проект по специализации Дата-инженер

Головня Алексей Юрьевич
г.Минск

Задача: наполнение БД новыми данными





Часть 1.

Первичное наполнение базы данных

Обработка датафрейма Sessions

- преобразование колонки `visit_date` в тип `datetime`

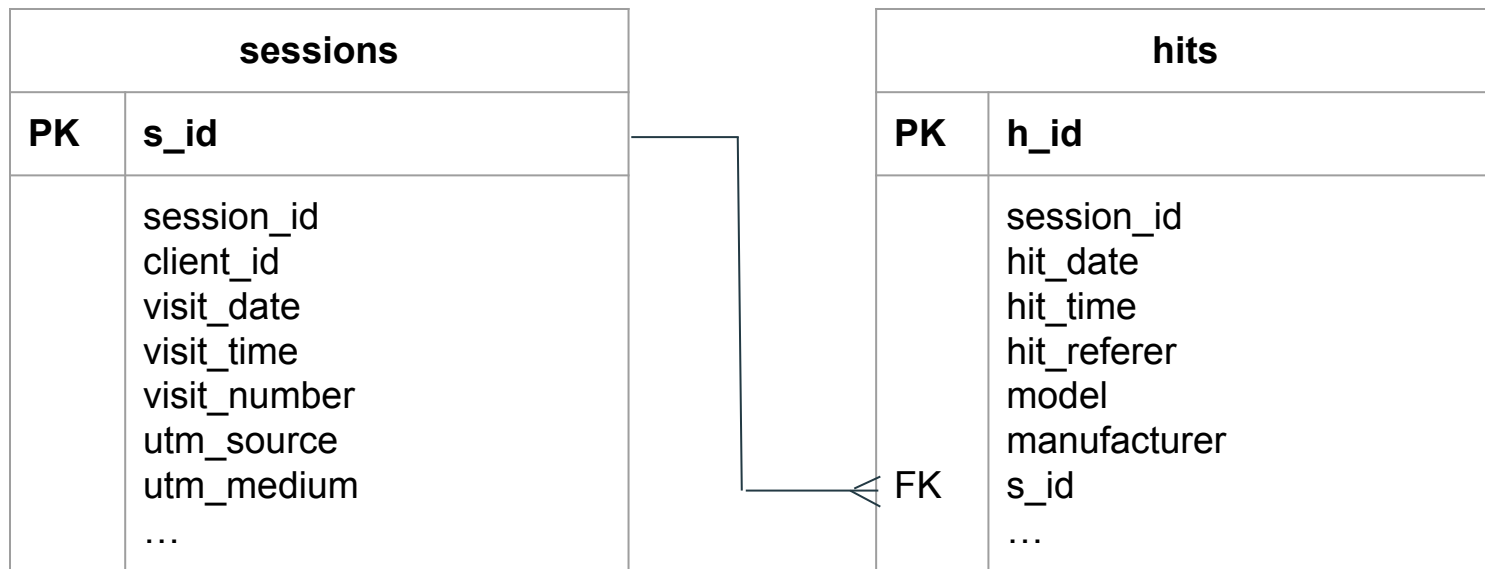
Обработка датафрейма Hits



- создание новых колонок 'manufacturer' и 'model' на основе колонки 'hit_page_path'
- удаление неинформативных колонок 'hit_type', 'event_value' и 'hit_page_path'
- преобразование колонки 'hit_date' в тип datetime

Перенос датафреймов в базу данных

- перенос датафрейма Sessions в БД
- создание primary key для таблицы БД Sessions
- добавление колонки 's_id' в датафрейм Hits
- перенос датафрейма Hits в БД
- создание primary key для таблицы БД Hits

Получившаяся база данных

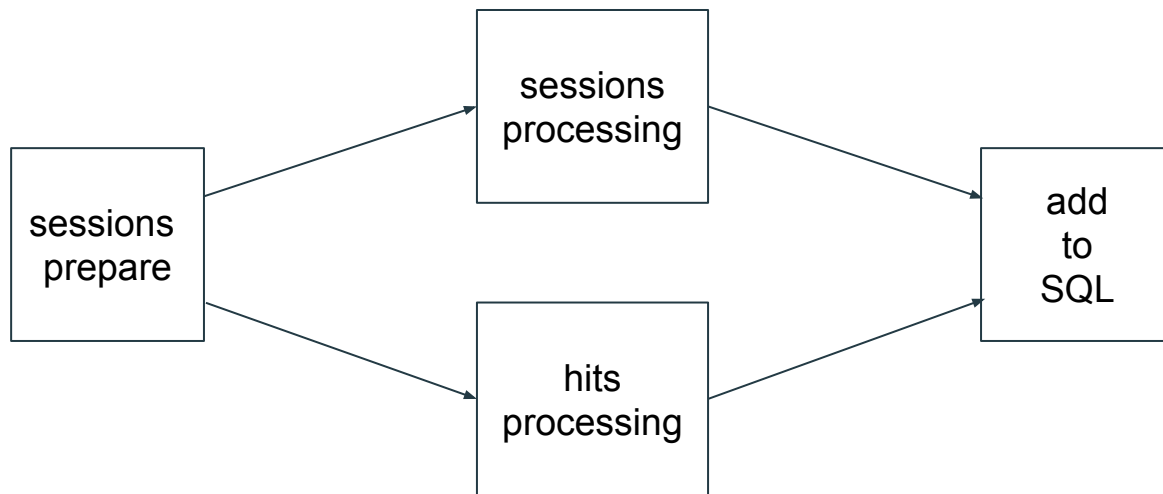




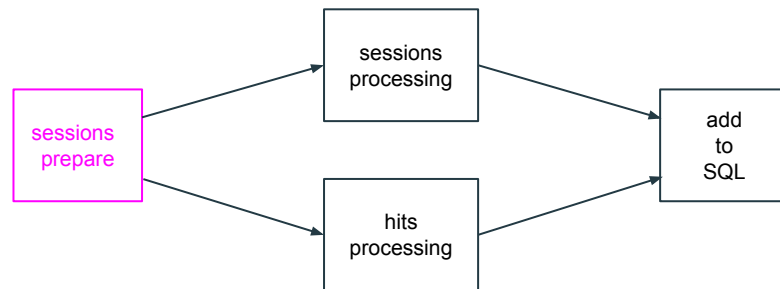
Часть 2.

Заполнение базы данных

Схема процесса



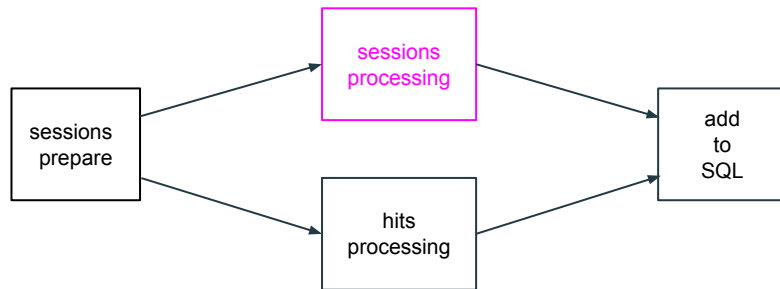
Подготовка данных Sessions



- открытие json файлов Sessions
- конвертация json в DataFrame
- сохранение DataFrame в pickle
- перенос обработанных файлов в папки 'processed+' и 'processed-'

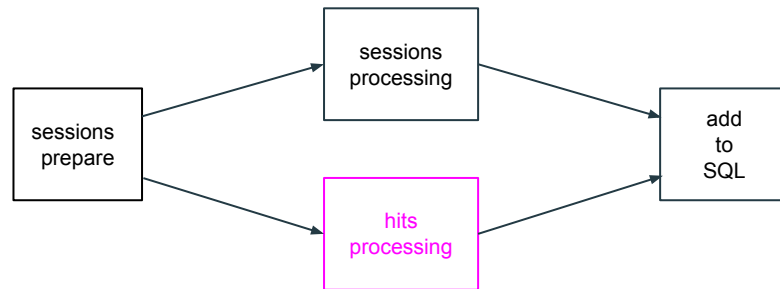
Обработка данных Sessions

- открытие файла pickle
- перенос данных в БД
- добавление в DataFrame колонки s_id из БД
- сохранение обновлённого DataFrame в pickle

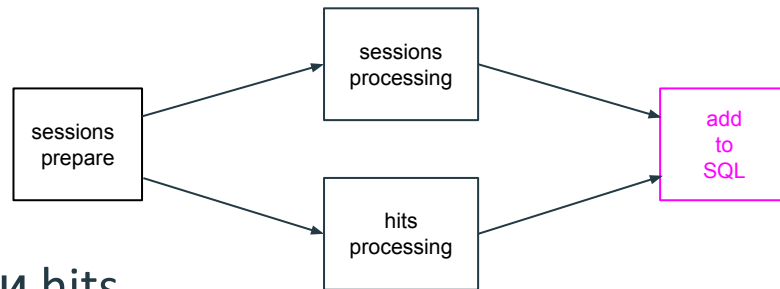


Обработка данных Hits



- открытие json файлов Hits
- конвертация json в DataFrame
- обработка DataFrame
- сохранение DataFrame в pickle



Перенос данных в БД



- открытие pickle файлов sessions и hits
- добавление в DataFrame hits колонки s_id
- перенос данных hits в БД



Часть 3.

Демонстрация работы Airflow.

Итоги



Спасибо за внимание!