Система сбора и хранения данных о просмотрах каналов (роликов) Youtube

# Цели и задачи

Цель системы состоит в регулярном сборе и хранение значимых метрик роликов и каналов Youtube с обеспечением доступа к этой информации с помощью систем визуализации (powerBI) или других систем.

Основной потребитель данных – аналитики, маркетинг.

# Технологический стек системы

Python 3.10, PostgreSQL, ~~Selenium~~

(отладка и разработка jupyter notebook)

# Структура и краткое описание системы

Для хранения данных используется БД PostgreSQL. Она состоит из 4х таблиц. Текущее состояние «частично» нормализованная по 2ой форме. В таблицах хранится:

1. список youtube каналов (yt\_channels)
2. список и описание youtube роликов (yt\_reels)
3. принадлежность роликов к каналам (yt\_channels2reels)
4. статистика по роликам (yt\_reels\_stat).

Выгрузка данных о просмотрах осуществляется с сайта popsters.ru (платная подписка) скриптом get\_popsters\_sql.py, который получает с сайта xlsx файлы с данными по роликам на канале. Список каналов и ссылки на них скрипт получает из таблицы БД. Также скрипт генерит два файла - report\_urls.csv – файл, в котором описывается привязка xlsx файлов и каналов, по которым они содержат информаицю и файл failed\_yt.csv – список каналов, по которым статистика (по каким-либо причинам) не была получена (выгрузка данных не произошла).

Вторым скриптом – yt\_parser\_loader.py – производится парсинг файла report\_urls.csv и последовательная обработка xlsx файлов. В ходе которой производятся следующие операции:

1. получение списка роликов и обновление таблицы роликов
2. получение списка роликов на данном канале и обновление таблицы yt\_channels2reels.
3. Обновление таблицы со статистикой по роликам yt\_reels\_stat

После обработки всех xlsx файлов (всех данных собранных по каналам) производится обновление таблицы yt\_reels\_stat – дублирование последних данных статистики на данную дату сбора информации и разметка удаленных роликов.

# Структура системы и БД (ERD)

