**Задача:** необходимо построить платформу для анализа данных с разных источников рекламы на сайт.

**Необходимо:** запускать периодически парсеры сбора метрик для будущего анализа.

**Желательно:** использовать простые инструменты, которые можно запускать через Python из Docker, использовать максимально готовые решения, которые можно быстро вывести в прод.

Для подобных решений лучше всего подходит платформа, которая периодически может запускать питоновские скрипты через обычный Python и легко устанавливается в переменное окружение как пакет – pip install airflow.

Пишется парсер -> запускается в платформе airflow - она имеет web-интерфейс включающий в себя:

* Запись логов запускаемых тасков
* Выдачу доступов различным пользователям
* Встроенной статистикой с триггерами и scheduler-ом

Далее необходимо автоматизировать деплой кода на серверах (Dokku/CI), которые интегрированы с gitlab, настроить хранение данных в сыром виде в BigQuery, и написать алгоритмы фильтрации и агргегации полученных данных в самом BigQuery.

Итого:

Web => Парсер (через Airflow набор питоновских кодов) => Все преобразуется в список словарей => через библиотеку airflow (from airflow.contrib.operators.bigquery\_operator import BigQueryOperator ) => BigQuery далее на StandartSQL мы можем создавать:

Таблицы фактов, собирать статистики и делать агрегаты типа:

select

spt, day, y, count(\*) as total

from (

SELECT

clientid, visitStartTime, TIMESTAMP(date) as day

SPLIT(hits.eventInfo.eventLabel "?") as spt

FROM TABLE\_DATE\_RANGE ([api-project.ga\_sessions\_] TIMESTAMP('2020-01-01') TIMESTAMP ('2020-02-10'))

where hits.eventInfo.eventLabel like '%test1%'

)

where spt like '%test2%'

group by 1, 2, 3

having total > 5 –условие на метрику

order by 1 desc, 2 desc

По нужным метрикам и событиям можно смотреть статистику через GoogleAnalytics.

Используемые ссылки на примеры:

<https://www.applydatascience.com/airflow/bigquery-pipeline-airflow/>

<https://cloud.google.com/blog/products/bigquery/how-to-aggregate-data-for-bigquery-using-apache-airflow>

<https://hoffa.medium.com/hacker-news-on-bigquery-now-with-daily-updates-so-what-are-the-top-domains-963d3c68b2e2>