

Airflow для маркетологов

Пошагово, без лишней информации и пояснений, что надо сделать

если нужен пример Дага, то пишите мне в ЛС или посмотрите код тут https://airflow-analytics.sbertech.io/tree?dag_id=main_marketing_dash

КАК НАПИСАТЬ СВОЙ ПЕРВЫЙ ДАГ

• Этап 1

1. написать в SD на получение доступов к аналитическому AirFlow
2. [создать заявку на доступ в Глобал протект](#) - наш ВПН (ниже инструкция, как его завести)



Инструкция по ис...е Access VPN.pdf

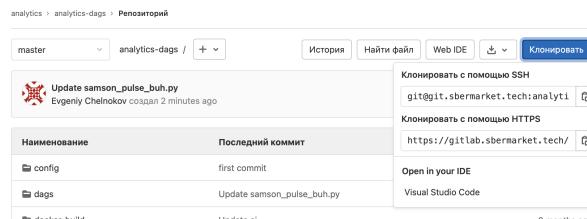
1. включить ВПН и зайти в наш эирфлоу

<https://airflow-analytics.sbertech.io/home>

• Этап 2

1. зайти на гитлаб в наш эирфлоу <https://gitlab.sbermarket.tech/analytics/analytics-dags/-/tree/master>
2. проверить что у вас есть ключ SSH (иконка профиля->Ключи SSH-> получить ключ-> пишете свой ключ) этот ключ, который вы создали себе сами, является паролем для входа в гит через терминал

3. открыть репозиторий <https://gitlab.sbermarket.tech/analytics/analytics-dags/-/tree/master> и склонировать его (попробуйте сначала с помощью SSH)



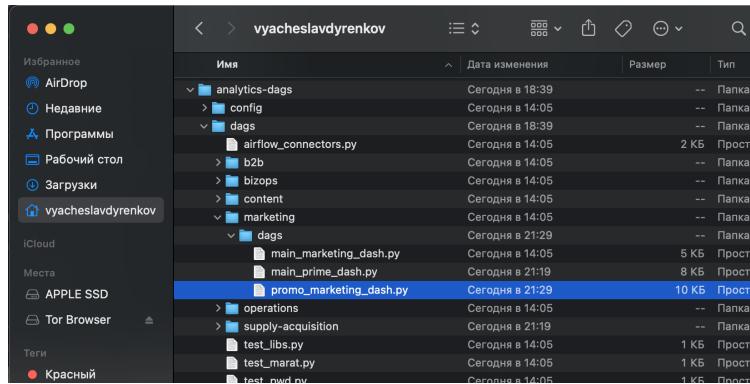
4. открыть терминал на компьютере и написать следующее

- a. **git init**
 - b. **git clone** (вставить через command+v то, что скопировано в пункте 3)
5. вылезет окошко с аутентификацией
- a. вводим свой логин (имя.фамилия)
 - b. вводим пароль (ключ SSH)

- конец этапа 2 -

• Этап 3

1. ищем в finder папку с названием **analytics-dags**, чтобы убедиться в ее существовании!!!
2. далее открываем терминал и переходим через команду **cd analytics-dags/dags/marketing/dags/** в папку **marketing** с нашими дагами



3. пишем команды в гитлабе следующем порядке

- a. **git checkout master**
 - b. **git pull** (общими словами получили мастер ветку с гитлаба)
 - c. **git branch -D DATA-3228** (удалили ветку с названием DATA-3228)
 - d. **git checkout -b DATA-3228** (создали ветку DATA-3228 нужно писать именно DATA-[номер таски] чтобы ваши кометы были видны в джире)
4. на этом этапе мы заходим через finder в папку **marketing/dags** (как на скрине) и вставляем файл или вносим изменения в код существующего файла, затем сохраняем через command+s
5. возвращаемся в терминал и вводим команду **git status**, должны оранжевым подсвечиваться файлы, которые вы изменили или добавили (примерно как на скрине, только там будет .py файл)

```
(используйте «git add <файл>...», чтобы добавить в то, что будет включено в коммит)
./.../.DS_Store
./.../.DS_Store

ничего не добавлено в коммит, но есть неотслеживаемые файлы (используйте «git add», чтобы их)
vyacheslavdyrenkov@MacBook-Pro-Vyacheslav_dags ~
```

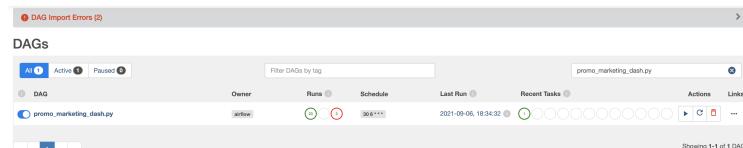
6. теперь пишем следующие команды

- a. **git add** [название файла который вы изменили/добавили].py (например git add promo_marketing_dash.py)
- b. **git commit -m "Комментарий к коммиту"** (обязательно пишите -m, иначе вас закинет в vim пример git commit -m "DATA-3228: добавил маму на авито"
- c. **git push --set-upstream origin DATA-3228** (DATA-3228 это уже название ветки, которую создали в пункте 3)
- d. далее, если все ок, у вас в терминале появится ссылка на merge request, копируем ее и вставляем в браузер, нажимаем слияние

- конец этапа 3 -

• Этап 4

1. проверяем <https://gitlab.sbermarket.tech/analytics/analytics-dags/-/tree/master/dags/marketing/dags> вот тут, что наши изменения прошли (или добавились новые файлы)
2. заходим в airflow в <https://airflow-analytics.sbmkt.io/home> и смотрим на самый верх



ВОТ ТУТ НЕ ДОЛЖНО БЫТЬ НИКАКИХ ОШИБОК, ЕСЛИ ВЫ ВСЕ СДЕЛАЛИ ПРАВИЛЬНО

1. заходим в поиск и вводим название нашего Дага (иногда нужно подождать 2-3 минутки пока он прогрузится в эирфлоу)

1. переходим по поиску в даг и включаем тумблер

All 1 Active 1 Paused 0

Filter DAGs by tag: promo_marketing_dash.py

DAG	Owner	Runs	Schedule	Last Run	Recent Tasks	Actions	Links
promo_marketing_dash.py	airflow	23 3	30 6 * * *	2021-09-06, 18:34:32	1	Trigger DAG Trigger DAG w/ config Edit Delete ...	View DAG

Showing 1-1 of 1 DAGs

DAG: [promo_marketing_dash.py](#) load promo_marketing_dash schedule: 30 6 * * *

Tree View Graph View Calendar View Task Duration Task Tries Landing Times Gantt Details Code

2021-09-06T18:34:32Z Runs 25 Update

PythonOperator

Auto-refresh [Trigger DAG](#) [Trigger DAG w/ config](#)

Aug 26, 04:30 Aug 27, 10:37 Aug 30, 04:30 Sep 02, 04:30 Sep 05, 06:30 Sep 06, 06:30

[DAG] main

1. нажимаем

<>code (самая последняя опция) и проверяем, что там код, который нам нужен (если нет, то ждем еще пару минуток)

1. нажимаем trigger dag и ждем

DAG: [promo_marketing_dash.py](#) load promo_marketing_dash schedule: 30 6 * * *

Tree View Graph View Calendar View Task Duration Task Tries Landing Times Gantt Details Code

2021-09-06T18:34:32Z Runs 25 Update

PythonOperator

Trigger DAG Trigger DAG w/ config Auto-refresh [Trigger DAG](#) [Trigger DAG w/ config](#)

Aug 26, 04:30 Aug 27, 10:37 Aug 30, 04:30 Sep 02, 04:30 Sep 05, 06:30 Sep 06, 06:30

[DAG] main

1. смотрим на цвет квадратика и нажимаем на него и переходим в log, чтобы понять есть ли ошибка или все ок.



8) если все ок (написано succes или нет ошибок), то супер, он работает

1. если есть ошибка, то останавливаем dag нажимаем на квадратик Clear ok

Task Instance: main

X

at: 2021-09-06T18:34:32.129926+00:00

Instance Details

Rendered

K8s Pod Spec

Log

All Instances

Filter Upstream

Download Log (by attempts):

1

Task Actions

Ignore All Deps

Ignore Task State

Ignore Task Deps

Run

Past

Future

Upstream

Downstream

Recursive

Failed

Clear

Cle
inst
sch

Past

Future

Upstream

Downstream

Mark Failed

Past

Future

Upstream

Downstream

Mark Success

Close

10) потом возвращаемся к этапам 2-3, чтобы исправить ошибки

КАК ЧЕРЕЗ AIRFLOW ОБНОВЛЯТЬ TABLEAU

- Этап 1

берем за шаблон даг https://airflow-analytics.sbmto.io/tree?dag_id=main_marketing_dash.py
не забываем **импортировать** import tableauclient as TSC!!!!

Далее по коду
тут есть 2 джобы

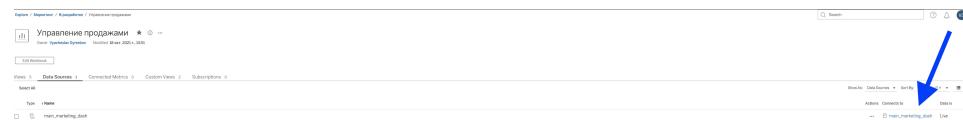
1. main() создает и апдейтит витрину
2. refresh_extract_tableau() обновляет экстракт

для каждой джобы прокинут свой коннекшн, откуда тянутся кредиты

```
clickhouse_connection = BaseHook.get_connection("clickhouse_analytics")
tableau_connection = BaseHook.get_connection("marketing_analytics_tableau")
```

• Этап 2

1. копируем себе даг из 1 пункта и в переменную TABLEAU_EXTRACT_NAME записываем название экстракта, как в табло.
Скрин с примером



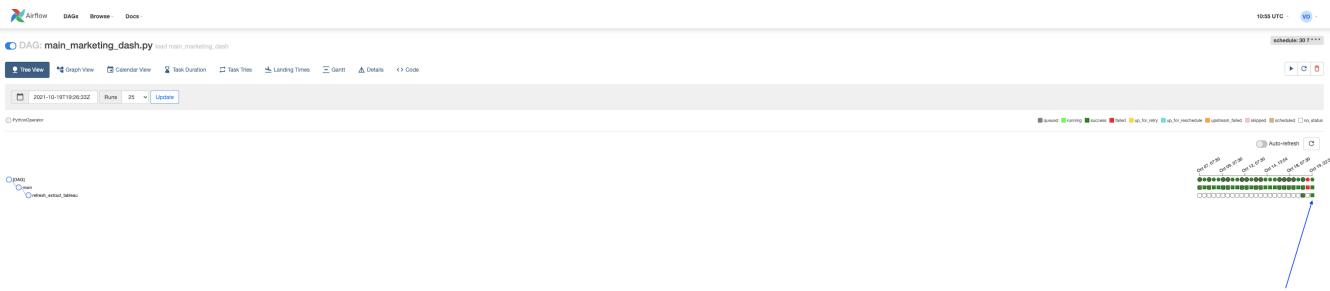
2. в самом конце дага будет прописано 2 джобы

```
with DAG('main_marketing_dash.py',
         description='load main_marketing_dash',
         schedule_interval='30 7 * * *',
         start_date=datetime(2020, 11, 1),
         catchup=False,
         max_active_runs=1,
         concurrency=1
     ) as dag:
    main = PythonOperator(
        task_id='main',
        python_callable=main
    )
    refresh_extract_tableau = PythonOperator(
        task_id='refresh_extract_tableau',
        python_callable=refresh_extract_tableau,
        op_kwargs = {'data_source_name': TABLEAU_EXTRACT_NAME}
    )
    main >> refresh_extract_tableau
```

нужно обратить внимание на места, на которые указывают стрелочки, они должны быть заполнены в соответствии с **ВАШИМИ** названиями **ФУНКЦИЙ** и **ВАШИМИ** названиями **ДАГОВ**

• Этап 3

Запускаем даг и смотрим, чтобы все отработало, то есть загорелись 2 зеленых квадратика



Первый квадратик должен вернуть 'success', означающий обновление витрины, второй квадратик отрабатывает мгновенно, так как он лишь **ЗАПУСКАЕТ** обновление экстракта, то есть он **не будет** гореть на протяжении всего обновления экстракта в табло.

P.S. Так как креды сделаны для **маркетинговой аналитики**, то есть вероятность, что может не завестись, пока owner экстракта не **Андрей Быстрков**, так что, если вы уверены, что все сделали правильно, попробуйте изменить владельца, как на скрине:

P.S.S. Экстракт должен быть **опубликован**

ПССС Положите в tags DAGa – `tableau-extract`, чтобы коллеги из BI команды могли узнать, кто дергает экстракты из AirFlow.

Ниже мем

