Руководство Администратора ETL
Системы бизнес-анализа BI (СБА)

10-01-2023

Сокращение	Описание
БД	База данных
ЕХД	Единое хранилище данных
Dag, даг	Directed Acyclic Graph это основной элемент Airflow, собирающий задачи вместе, организованный с помощью зависимостей и взаимосвязей, чтобы указать, как они должны выполняться.
ETL	трёхэтапный процесс управления данными, который означает «извлечение, преобразование, загрузка
ПО	Программное обеспечение

Перечень используемых терминов	Описание
Администратор	Работник, должностные обязанности которого заключаются в обеспечении штатной работы Системы, проведение аварийных работ.

Основной функционал ЕХД.

Целевая аудитория.

Настоящее руководство администратора предназначено для пользователей программного обеспечения – программы ЭВМ – «Система бизнес-анализа BI» (далее – СБА).

Назначение документа.

Документ содержит описание последовательности действий в основных операциях, которые должны производить администраторы СБА.

Необходимая подготовка специалиста по администрированию СБА.

Основные компетенции Администратора СБА:

- навыки работы с ОС Linux на уровне продвинутого пользователя;

- навыки работы с Docker;

- навыки работы с БД Postgres на уровне продвинутого пользователя;

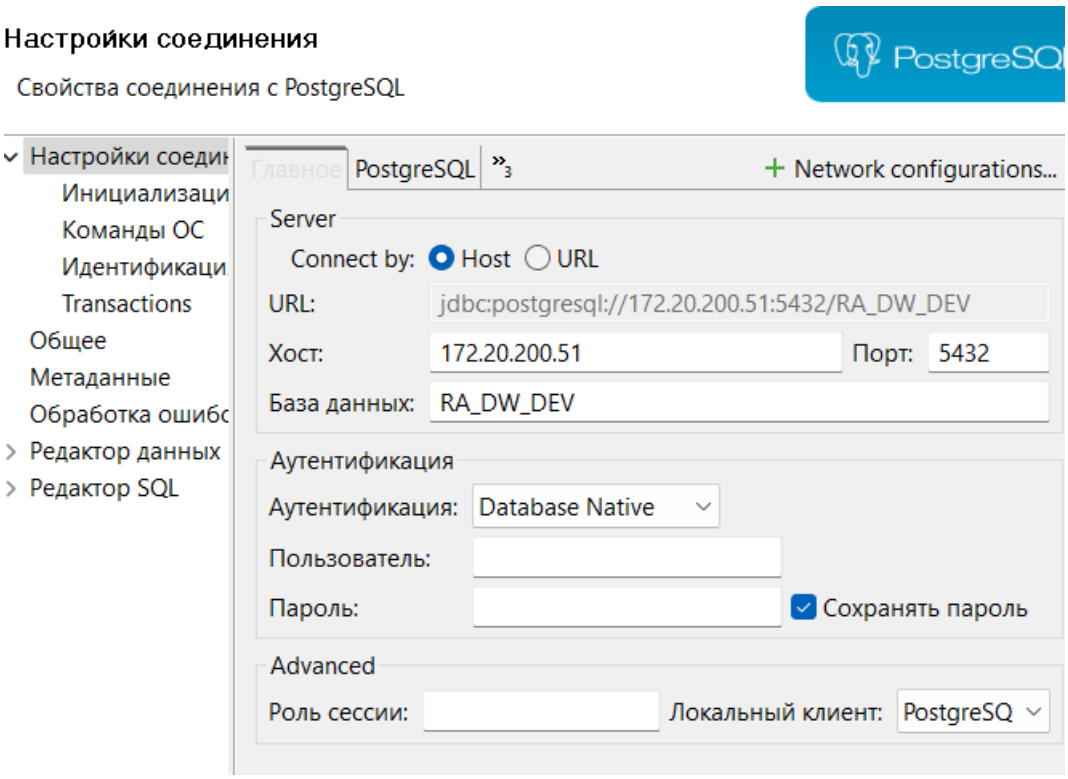
- навыки работы с Apache Airflow на уровне продвинутого пользователя.

Авторизация СБА

1) Авторизация в БД

Для входа в ЕХД СБА необходимо создать подключение к БД через SQL-клиент.

Параметры подключения к БД:



Учетная запись предоставляется по требованию.

2) Авторизация в интерфейсе управления процессами загрузки и обработки данных Airflow.

Для входа в интерфейс управления процессами необходимо:

1) В браузере перейти по адресу: http://172.20.200.41:8080/login/
2) Учетные данные Администратору предоставляются дополнительно.

Описание установки и настройки инструментов.

Реализация процедур ETL реализована с использованием ПО Apache Airflow v.2.4.

Фреймворк Apache Airflow развернут на виртуальной машине под управлением операционной системы Ubuntu Server. Установка ПО сделана с использованием Docker контейнеризации.

Apache Airflow развернут на машине *172.20.200.41* вход на сервер осуществляется через TTY консоль, SSH протокол. Пользователь airflow.

Для работы с контейнерами Airflow необходимо перейти в папку установка ПО

cd /opt/airflow

```
airflow@ecs-dwh-vm1:/opt/airflow$ ls -l
total 100
-rw-r--r--  1 root root 49634 Oct  3 21:56 airflow.cfg
drwxrwxr-x  7 root root  4096 Jan 22 12:21 dags
-rw-rw-r--  1 root root 10739 Jan 18 23:41 docker-compose.yaml
-rw-r--r--  1 root root 10732 Jan 11 10:07 docker-compose.yaml.bak
-rw-r--r--  1 root root   150 Jan 18 23:39 Dockerfile
drwxr-xr-x  2 root root  4096 Oct 27 22:49 libs
drwxrwxr-x 20 root root  4096 Jan 22 12:38 logs
drwxrwxr-x  2 root root  4096 Sep 25 23:07 plugins
drwxr-xr-x  2 root root  4096 Jan 30 03:00 tmp_files
airflow@ecs-dwh-vm1:/opt/airflow$
```

Проверка статуса сервисов.

docker container ls

```
airflow@ecs-dwh-vm1:/opt/airflow$ docker container ls
CONTAINER ID   IMAGE                      COMMAND                  CREATED      STATUS                    PORTS                                       NAMES
e672e2e29299   apache/ra_airflow:1.0.1    "/usr/bin/dumb-init …"   6 days ago   Up 6 days (unhealthy)     8080/tcp                                    airflow-airf
low-triggerer-1
0b7c1f072ef8   apache/ra_airflow:1.0.1    "/usr/bin/dumb-init …"   6 days ago   Up 6 days (healthy)       0.0.0.0:8080->8080/tcp, :::8080->8080/tcp   airflow-airf
low-webserver-1
89b662567f06   apache/ra_airflow:1.0.1    "/usr/bin/dumb-init …"   6 days ago   Up 6 days (unhealthy)     8080/tcp                                    airflow-airf
low-worker-1
e8fb68ad3f6a   apache/ra_airflow:1.0.1    "/usr/bin/dumb-init …"   6 days ago   Up 6 days (unhealthy)     8080/tcp                                    airflow-airf
low-scheduler-1
116c742bf287   postgres:13                "docker-entrypoint.s…"   6 days ago   Up 6 days (healthy)       5432/tcp                                    airflow-post
gres-1
1e73124342f5   redis:latest               "docker-entrypoint.s…"   6 days ago   Up 6 days (healthy)       6379/tcp                                    airflow-redi
s-1
```

Запуск Apache Airflow

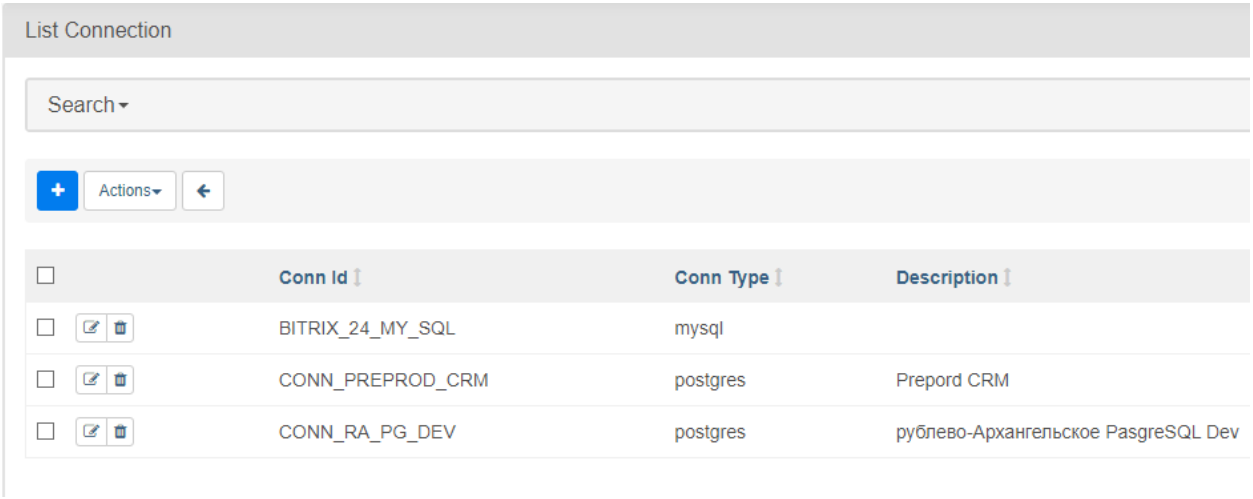
Docker compose up

Остановка Apache Airflow

Docker compose down

WEB интерфейс Apache Airflow доступен через браузер, по адресу http://172.20.200.41:8080/home

В Apache Airflow преднастроены соединения к используемым БД.



Так же установлен ряд служебных переменных (Variables).

Для загрузки данных написан ряд Airflow DAGs, код DAGs расположен в /opt/airflow/dags.

Переиспользуемый в DAGs код вынесен в отдельные python модули /opt/airflow/dags/lib.

Конфигурационные параметры для загрузки данных расположены в yaml файлах.

/opt/airflow/dags/config

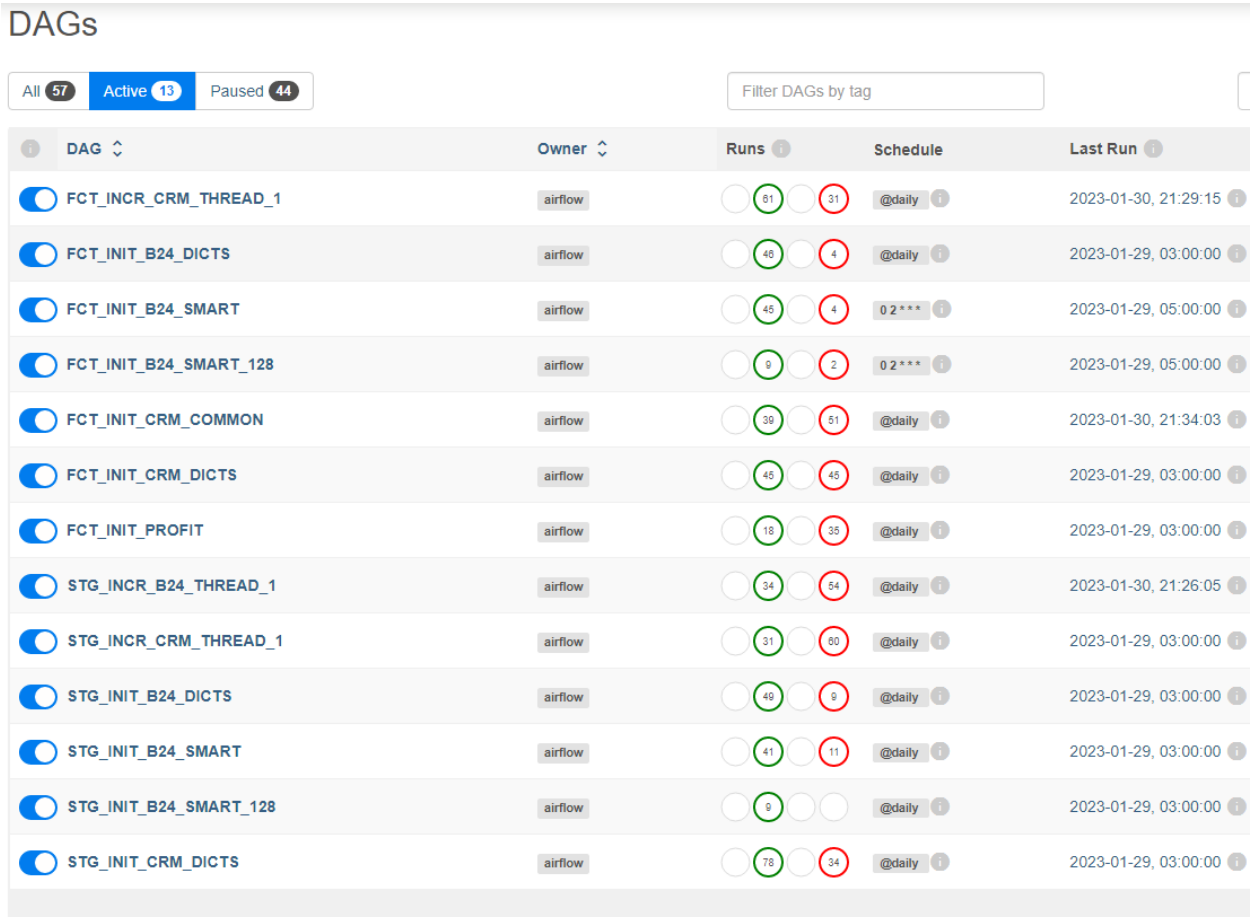
Конфигурация содержит данные об источнике, целевой таблице, названия целевой схемы, [перечень целевых столбцов], флаг необходимости очистки

таблицы перед вставкой, словарь для подстановок source-target (преобразования типов, применение функций и т.д.).

Пример конфигурации для ETL

```
b24_deal:
  source: DWH
  src_table: stage.b24_deal
  dest_schema: facts
  dest_fields:
  dest_truncate_yn: n
  substs:
    ddu_payment_method: ddu_payment_method::int
```

Процессы загрузки запускаются автоматически по расписанию начиная с 0:00 (UTC)



Результаты работы DAGs доступны через веб-интерфейс. Доступны как общая информация о статусе загрузки, так и детальная информация (logs)

Более подробную информацию о работе с Apache Airflow можно получить в официальной документации к проекту https://airflow.apache.org/docs/

Администрирование пользователей СБА.

Создание пользователя:

Для создания нового пользователя в Системе требуется войти в БД под учетной записью Администратора и выполнить скрипт создания пользователя указав имя и пароль. (пример маски имени пользователя: Фамилия_ИО/Belov_EA)

CREATE USER username WITH PASSWORD '****';**

Редактировании пользователя:

Для редактирования пользователя требуется войти в БД и выполнить скрипт указав имя пользователя и прописав дополнительные модификаторы в поле «OPTION».

ALTER USER username WITH [OPTION];

Блокировка пользователя:

Для блокировки пользователя требуется войти в БД и выполнить скрипт, указав имя пользователя.

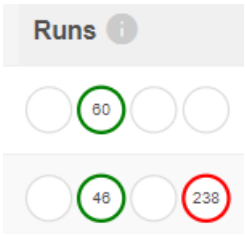
ALTER USER username NOLOGON;

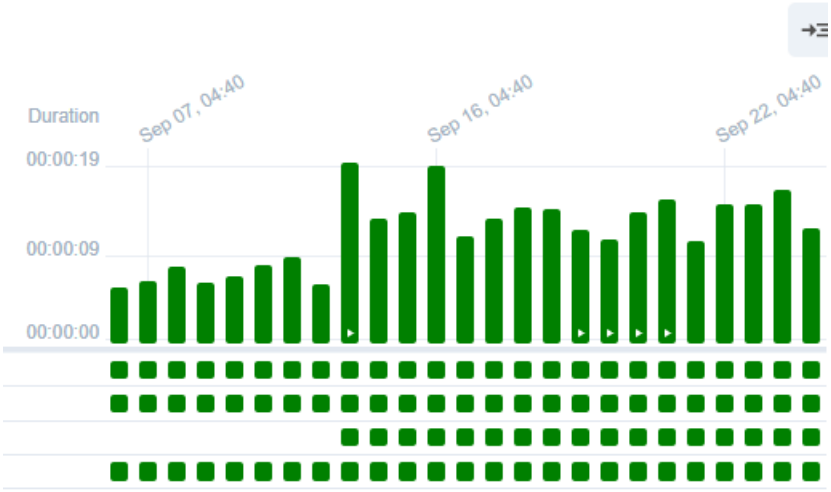
Администрирование процессов загрузки и обработки данных Airflow

Включение/выключение процесса.

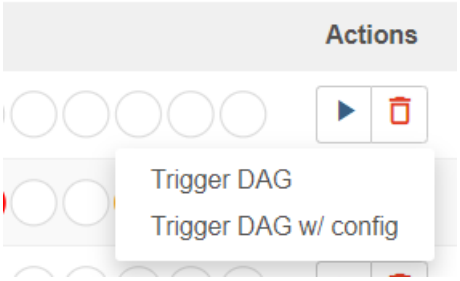
Для работы в Системе необходимо авторизоваться в интерфейсе. В главном меню нажать кнопку DAG. Нажать на кнопку:

Контроль успешности запусков DAG осуществляется через поле RUNS

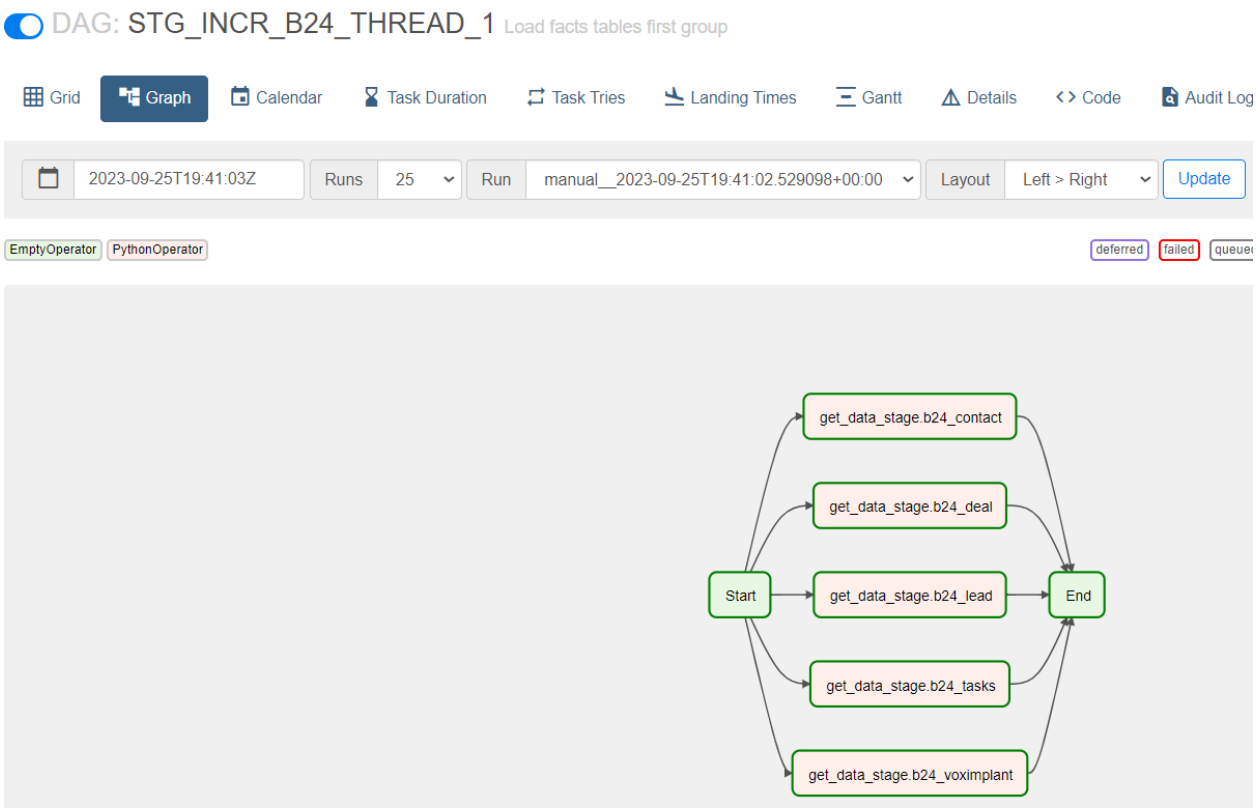




Для ручного старта отработки DAG, необходимо нажать на стрелочку и выбрать Trigger DAG



Для понимания структуры DAG, можно перейти на следующее окно.



На главном меню можно видеть активные/остановленные DAG.

Дату и время последнего/следующего запуска.