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 | Описание |
|-----------------------|------------------------------------|
| терминов | |
| 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 ВІ» (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 OC 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 SQLклиент.

Параметры подключения к БД:

| | Настройки | соединения |
|--|-----------|------------|
|--|-----------|------------|

| Настройки соеди | | 🖗 PostgreS | | | | |
|---|--|-------------------------|---------------------------------|--------------------|--|--|
| Свойства соединени | ия с PostgreSQL | | | | | |
| Настройки соедин Инициализаци Команды ОС Идентификаци. | Главное Postgre Server Connect by: | eSQL N³ ● Host ○ URL | + Netw | vork configuration | | |
| Transactions Общее Метаданные Обработка ошибс | URL: | jdbc:postgresql://172. | 20.200.5 <mark>1:</mark> 5432/F | 432/RA_DW_DEV | | |
| | Хост: | 172.20.200.51 | | Порт: 5432 | | |
| | База данных: | | | | | |
| > Редактор данных | Аутентификац | ия | | | | |
| > Редактор SQL | Аутентификац | ия: Database Native | ~ | | | |
| | Пользователь | : | | | | |
| | Пароль: | | 🔽 Co | охранять пароль | | |
| | Advanced | | | | | |
| | Роль сессии: | | Локальный кли | ент: PostgreSQ > | | |

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 | 5-di | wh-vmi | L:/op | t/airf | low\$ | ls | - | | | |
|-------------|------|--------|-------|--------|-------|-----------|-------|-------------------------|--|--|
| total 100 | | | | | | | | | | |
| -rw-rr | 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-rr | 1 | root | root | 10732 | Jan | 11 | 10:07 | docker-compose.yaml.bak | | |
| -rw-rr | 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 | plūgins | | |
| drwxr-xr-x | 2 | root | root | 4096 | Jan | 30 | 03:00 | tmp_files | | |
| airflow@ecs | s-di | wh-vm: | L:/op | t/airf | low\$ | | | | | |

Проверка статуса сервисов.

docker container ls

| airflow@ecs-dw | n-vml:/opt/airflow\$ docker | container Is | | | | |
|----------------|-----------------------------|------------------------|------------|-----------------------|---|--------------|
| 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- | 4 | | | | | |
| 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- | 4 | | | | | |
| 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- | 4 | | | | | |
| 116c742bf287 | postgres:13 | "docker-entrypoint.s" | 6 days ago | Up 6 days (healthy) | 5432/tcp | airflow-post |
| gres-1 | | | | | | |
| Ĩe73124342f5 | redis:latest | "docker-entrypoint.s" | 6 days ago | Up 6 days (healthy) | 6379/tcp | airflow-redi |
| 5-1 | | | | | | |

Запуск Apache Airflow Docker compose up

Остановка Apache Airflow

Docker compose down

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

В Apache Airflow преднастроены соединения к используемым БД.

| List Connection | | | | | | | | | |
|-----------------|------------------|-------------|-------------------------------------|--|--|--|--|--|--|
| Search - | | | | | | | | | |
| + Actions - ← | | | | | | | | | |
| | Conn Id I | Conn Type 🎗 | Description 1 | | | | | | |
| | BITRIX_24_MY_SQL | mysql | | | | | | | |
| | CONN_PREPROD_CRM | postgres | Prepord CRM | | | | | | |
| | CONN_RA_PG_DEV | postgres | рублево-Архангельское PasgreSQL Dev | | | | | | |

Так же установлен ряд служебных переменных (Variables).

| List Variable | | | |
|---------------------|-------------------|------------------------------------|--------------------------|
| Search - | | | |
| + Actions- ← | | | |
| | Key 1 | Val 1 | Description [|
| | B24 | { "CONNECTION_ID" : "BITRIX_2 | Bitrix24 |
| | CRM | { "CONNECTION_ID" : "CONN_P | Параметры источника CRM |
| | DWH | { "CONNECTION_ID" : "CONN_R | Параметры DWH |
| | DWH_CONNECTION_ID | CONN_RA_PG_DEV | DWH connection |
| | PROFIT_URL | https://pb12354.profitbase.ru/expo | Ссылка на отчет "Profit" |

Для загрузки данных написан ряд Airflow DAGs, код DAGs paсположен 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



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

| DAGs | | | |
|----------------------------|---------|---------------------|------------------------|
| All 57 Active 13 Paused 44 | | Filter DAGs by tag | |
| ● DAG ≎ | Owner 🗘 | Runs 🕕 Schedule | Last Run 🕕 |
| FCT_INCR_CRM_THREAD_1 | airflow | | 2023-01-30, 21:29:15 🕕 |
| | airflow | (4) (4) (adaily (1) | 2023-01-29, 03:00:00 🚺 |
| FCT_INIT_B24_SMART | airflow | | 2023-01-29, 05:00:00 🕕 |
| FCT_INIT_B24_SMART_128 | airflow | 9 2 02*** | 2023-01-29, 05:00:00 🕕 |
| | airflow | 39 (51) @daily [| 2023-01-30, 21:34:03 🕕 |
| | airflow | 45 45 @daily 1 | 2023-01-29, 03:00:00 🕕 |
| | airflow | 18 (35) @daily 1 | 2023-01-29, 03:00:00 🕕 |
| STG_INCR_B24_THREAD_1 | airflow | 34 54 @daily | 2023-01-30, 21:26:05 🕕 |
| STG_INCR_CRM_THREAD_1 | airflow | 31 (60) @daily [| 2023-01-29, 03:00:00 🕕 |
| | airflow | (49) (9) @daily (1 | 2023-01-29, 03:00:00 🕕 |
| STG_INIT_B24_SMART | airflow | (41) (11) @daily 1 | 2023-01-29, 03:00:00 🕕 |
| STG_INIT_B24_SMART_128 | airflow | G @daily | 2023-01-29, 03:00:00 🕕 |
| | airflow | (78) (34) @daily (1 | 2023-01-29, 03:00:00 🕕 |

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 <u>https://airflow.apache.org/docs/</u>

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_ИO/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.

DAG: STG_INCR_B24_THREAD_1 Load facts tables first group

| 🎛 Gri | d 📲 Graph | 🖬 Calend | lar 🎴 | Task Durati | on | 🛱 Task Tries | ▲ Landing Times | - Gantt | ▲ Details | <> Code | Audit Log |
|-----------|-----------------|----------|-------|-------------|-----|--------------|-----------------------|--|--|-------------|--------------|
| | 2023-09-25T19:4 | 11:03Z | Runs | 25 🗸 | Run | manual_202 | 3-09-25T19:41:02.529(| 098+00:00 ✔ | Layout L | eft > Right | ✓ Update |
| EmptyOper | PythonOperator | | | | | | | | | deferred | failed queue |
| | | | | | | | Start | get_data_stage. get_data_stage get_data_stage get_data_stage get_data_stage.b2 | b24_contact e. b24_deal e. b24_lead e. b24_lead | End | |

На главном меню можно видеть активные/остановленные DAG.

Дату и время последнего/следующего запуска.

| All 59 Active 14 Paused 45 | | | Filter DAGs by tag | | Search DAGs | | Auto-refresh | | |
|----------------------------|---------|---------|--------------------|------------------------|------------------------|--------------|--------------|---------|-------|
| ⑦ DAG ↓ | Owner 🗘 | Runs 🕚 | Schedule | Last Run 🕕 | Next Run 🗘 🕕 | Recent Tasks | | Actions | Links |
| FCT_INCR_DBT_MODELS_B24 | airflow | \odot | 40 4**** | 2023-09-25, 04:40:00 🚯 | 2023-09-26, 04:40:00 🛞 | 00000 | 00000000 | ÞŌ | |
| | airflow | 0 | 238) @daily | 2023-09-25, 00:00:00 🕕 | 2023-09-26, 00:00:00 🔘 | 00000 | | ÞŌ | |
| FCT_INIT_B24_SMART | airflow | | 20 4 **** | 2023-09-25, 04:20:00 🕕 | 2023-09-26, 04:20:00 🔘 | 00000 | 00000000 | | |
| FCT_INIT_B24_SMART_128 | airflow | 000 | 7 02*** | 2023-09-25, 02:00:00 🕕 | 2023-09-26, 02:00:00 🕕 | 00000 | 00000000 | • 0 | |
| FCT_INIT_CRM_B24_COMMON | airflow | | 30 4 *** 🚺 | 2023-09-25, 19:43:19 📵 | 2023-09-26, 04:30:00 🕚 | 00000 | 00000000 | ► Ō | |
| | | | | | | | | | |