Itera Data Platform

Инструкция по развертыванию

Оглавление

1 Аннотация

В настоящем документе представлено описание состава установочного пакета Itera Data Platform, а также подробное руководство по развертыванию системы.

В связи с высокой сложностью развертывания компонентов Itera Data Platform установку системы осуществляет вендор.

2 Установочный пакет

Развертывание системы выполняется с использованием установочного пакета itera.zip. Состав каталогов установочного пакета приведен в таблице.

Каталог	Описание
itera	Корневой каталог установочного пакета
dwh	Файлы установки хранилища данных
etl	Файлы установки модулей ETL
airflow	Файлы установки модуля управления конвейерами данных
dbt	Файлы установки модуля управления хранилищем данных

3 Описание контейнеров и служебных каталогов

Службы системы устанавливаются в виде отдельных контейнеров. В таблице приведен перечень разворачиваемых контейнеров служб, а также описание служебных каталогов.

Компонент	Контейнер	Каталог хоста	Каталог контейнера	Описание
Хранилище данных	postgres-dwh	/opt/dwh/postgres		Каталог размещения файлов конфигурации контейнера postgres-dwh.
		/opt/dwh/postgres/data	/var/lib/postgresql/data	Каталог файлов баз данных PostgreSQL.
		/opt/dwh/postgres/datafiles	/var/lib/postgresql/datafiles	Каталог для загрузки файлов данных. Настраивается общий NFS каталог для подключения с сервера ETL и обмена файлами данных между службами ETL и DWH.
		/opt/dwh/postgres/backup	/var/lib/postgresql/backup	Каталог сохранения дампов баз данных.
		/opt/dwh/postgres/scripts		Скрипты инициализации системных объектов и расширений базы данных хранилища.
Модуль управления хранилищем данных	dbt	/opt/etl/dbt		Каталог размещения файлов конфигурации контейнера dbt.
		/opt/etl/dbt/dbt_projects	/usr/app/dbt/dbt_projects	Каталог проектов dwh.
Модуль управления конвейерами данных	airflow-webserver airflow-scheduler airflow-worker-1 airflow-worker-2	/opt/etl/airflow		Каталог размещения файлов конфигурации контейнеров и файла конфигурации служб Airflow (airflow.cfg).
		/opt/etl/airflow/dags	/opt/airflow/dags	Библиотеки и файлы конфигурации конвейеров данных.
		/opt/etl/airflow/data	/opt/airflow/data	Каталог для размещения файлов данных, обрабатываемых конвейерами данных.
		/opt/etl/dbt	/opt/airflow/dbt	Каталог для получения скриптов моделей dwh, скомпилированных сервисом dbt. Каталог, используемый совместно с контейнером dbt.
		/opt/etl/airflow/logs	/opt/airflow/logs	Файлы журналов Airflow.

4 Инструкция по развертыванию

Хранилище данных может быть развернуто как на одном сервере с модулями ETL, так и на отдельном сервере.

Каталог установки на серверах развертывания – /opt. Пользователю, под которым осуществляется установка, необходимо предварительно предоставить права на чтение и запись к каталогу установки.

4.1 Развертывание хранилища данных

1. В файле переменных окружения itera/dwh/postgres/.env, при необходимости, скорректировать параметры:
   - POSTGRES_PASSWORD (пароль пользователя базы данных хранилища).
2. В itera/dwh/postgres/docker-compose.yml скорректировать параметры:
   - networks/subnet (указать диапазон IP-адресов, не конфликтующий с другими подсетями).
3. Скопировать содержимое каталога itera/dwh на сервер развертывания в /opt/dwh.
4. Запустить контейнер postgres-dwh:

```
sudo docker-compose -f /opt/dwh/postgres/docker-compose.yml up -d
```

4.2 Развертывание модуля управления хранилищем данных

1. В файл конфигурации профилей подключения к хранилищу данных itera/etl/dbt/dbt_projects/profiles.yml скорректировать параметры подключения:
   - host (указать IP или FDQN сервера хранилища данных),
   - password (пароль пользователя базы данных хранилища).
2. Скопировать itera/etl/dbt на сервер развертывания в /opt/etl/dbt.
3. Запустить контейнер dbt:

```
sudo docker-compose -f /opt/etl/dbt/docker-compose.yml up -d
```

4.3 Развертывание модуля управления конвейерами данных

1. В файле переменных окружения itera/etl/airflow/.env, при необходимости, скорректировать параметры:
   - POSTGRES_USER (логин пользователя базы данных airflow),
   - POSTGRES_PASSWORD (пароль пользователя базы данных airflow),
   - POSTGRES_HOST (указать IP или FDQN сервера хранилища данных),
   - AIRFLOW_USER_PASSWORD (пароль администратора airflow).
2. В itera/etl/airflow/docker-compose.yml скорректировать параметры:
   - networks/subnet (указать диапазон IP-адресов, не конфликтующий с другими подсетями).
3. Скопировать содержимое каталога itera/etl/airflow на сервер развертывания в /opt/etl/airflow.
4. Запустить контейнеры Airflow:

```
sudo docker-compose -f /opt/etl/airflow/docker-compose.yml up -d
```

5. Выполнить инициализацию основных подключений сервера ETL:

   5.1. В файле itera\etl\airflow\scripts\init_connections.sh, при необходимости, скорректировать команды инициализации подключений:

   - postgres_dwh - подключение к базе данных хранилища,
   - ssh_etl - подключение к серверу ETL по SSH,
   - smtp_default - SMTP подключение к серверу отправки сообщений.

   5.2. Выполнить на сервере развертывания скорректированные в п. 5.1 команды.

4.4 Настройка загрузки файловых источников данных

Для обеспечения возможности загрузки данных из файловых источников необходимо обеспечить для сервера базы данных хранилища доступность файлов данных, загружаемых модулем ETL. Для этого требуется связать каталог файлов данных на сервере ETL с каталогом файлов данных на сервере хранилища данных одним из двух способов:

1. При развертывании хранилища данных и модулей ETL на одном сервере создать в /opt/dwh/postgres/ символическую ссылку на каталог /opt/etl/airflow/data/datafiles:

```
sudo ln -s /opt/etl/airflow/data/datafiles /opt/dwh/postgres/datafiles
```

2. При развертывании хранилища данных и модулей ETL на разных серверах смонтировать каталог /opt/dwh/postgres/datafiles на сервере хранилища данных в каталог /opt/etl/airflow/data/datafiles на сервере ETL. После монтирования необходимо выполнить перезапуск контейнеров Airflow:

```
sudo docker-compose -f /opt/etl/airflow/docker-compose.yml restart
```