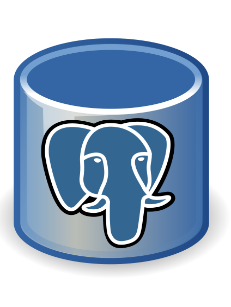
Архитектура решения

Схема архитектуры

Etl\_db\_4

ODS-слой

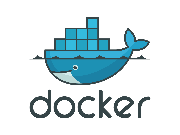
DM-слой

DDS-слой

intermediate-слой

Источник данных

Source



dashboards

Broken\_data



Слои обработки и управления

Описание компонентов

В приложении Docker поднят контейнер с Apache Airflow. Airflow обеспечивает оркестрацию ETL процесса. **Обработка данных происходит с помощью Python скриптов с использованием sql запросов и библиотеки pandas. Для визуализации используется** Luxms BI.

Источник данных:

**Исходные Данные хранятся в PostgreSQL в схеме source\_data.**

Слои обработки и управления:

1. ODS-слой: слой хранения данных из источника в «сыром», необработанном виде

Описание таблиц и атрибутов:

1. Intermediate-слой: слой хранения обработанных и приведённых к нужным форматам данных;
2. DDS-слой: слой хранения данных с созданными связями между таблицами и определенными ключевыми атрибутами;
3. DM-слой: слой специализированных витрин данных, сгруппированных в форме схемы «звезда».

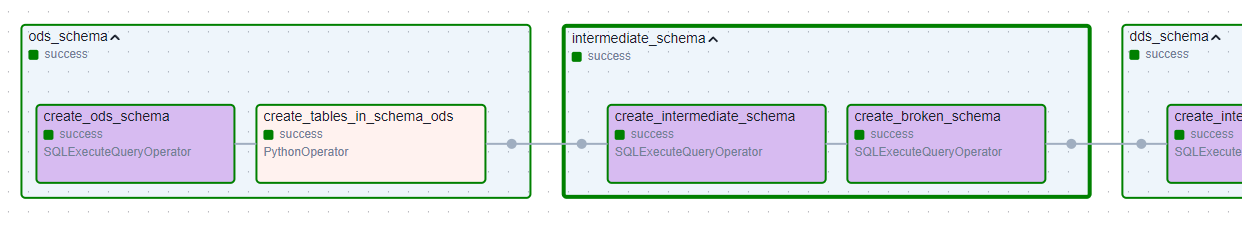
Дополнительно: схема broken\_data. Создана для хранения «битых» данных из источника. Заполняется в два этапа:

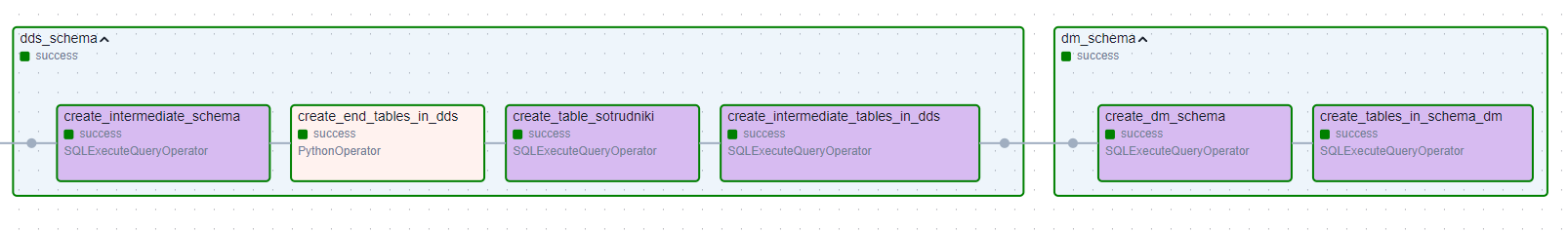
* во время формирования intermediate-слоя (если в записи есть пропуск по определенному атрибуту);
* во время формирования DDS-слоя (при отсутствии ключа в таблице «сотрудники\_дар» на который ссылается запись в промежуточной таблице)

Описание ETL потока

1. Создание схем в базе данных etl\_db\_4 и таблиц со связями происходит с помощью скрипта init\_schemes\_and\_tables.py. Скрипт запускается единожды и представляет собой даг, состоящий из следующих тасков:

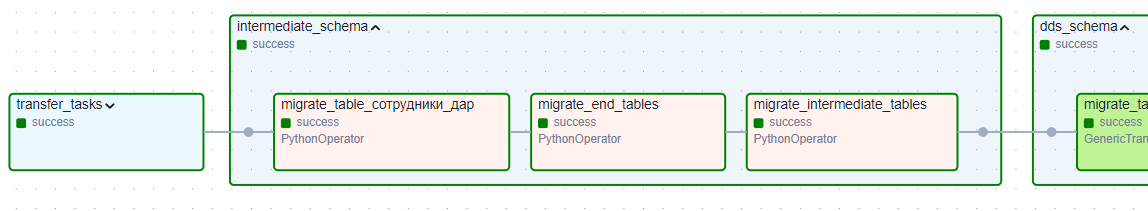
task\_group\_ods\_schema >> task\_group\_intermediate\_schema >> task\_group\_dds\_schema >> task\_group\_dm\_schema





Для каждой схемы созданы свои таскгруппы. В каждой таскгруппе, в зависимости от схемы, есть таски, отвечающие за создание схемы и создание таблиц.

1. Скрипт update\_data.py отвечает за последовательную загрузку данных в слои. Данный даг срабатывает раз в сутки.



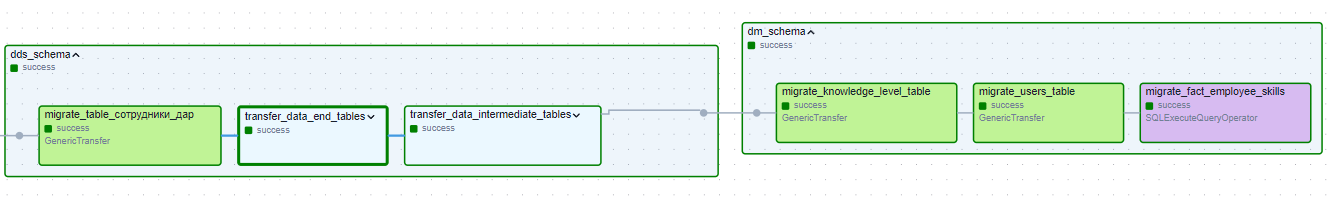


Схема витрины данных

