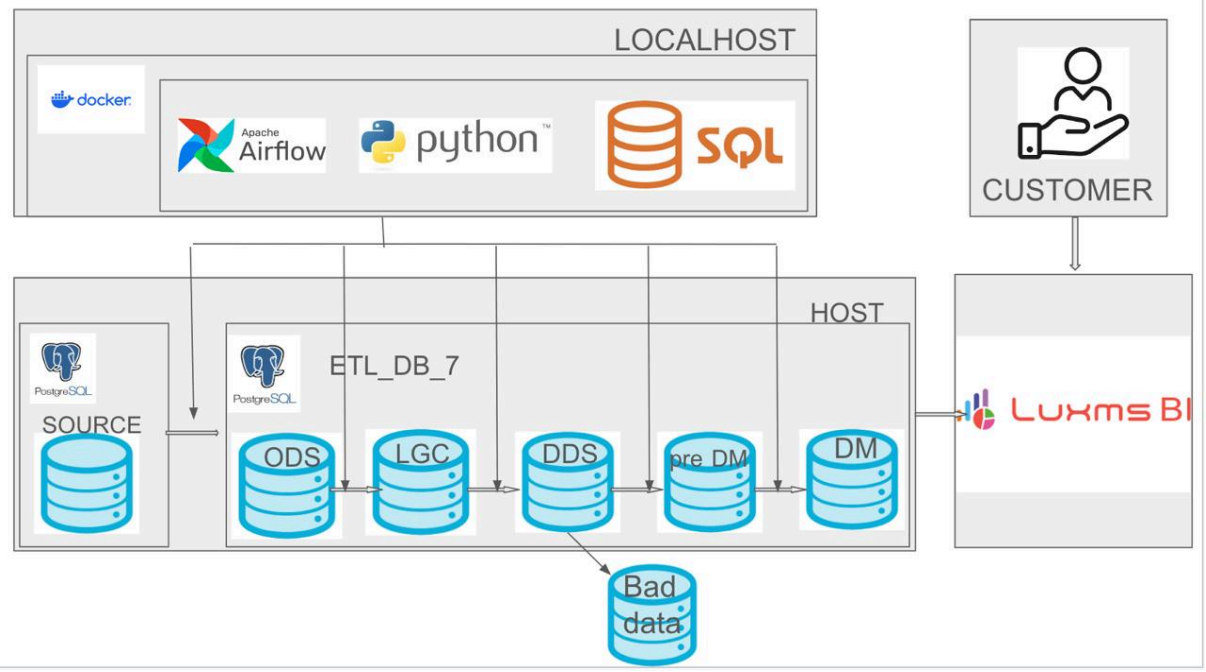
Архитектура решения

Стажировка в ДАР / Группа 7
Стажеры:
Геннадий Хазарьян
Мария Новожилова

ссылка на гугл-док: https://docs.google.com/document/d/1Iy6CLCYfA1ob-OCCY_NIy1kflD7nhAJWcIpHSVSm0JE/edit?usp=sharing

1. Схема архитектуры



2. Описание компонентов

Описание слоя управления:

В Docker поднят контейнер с Airflow. Airflow оркестрирует ETL-поток на базе python-скриптов.

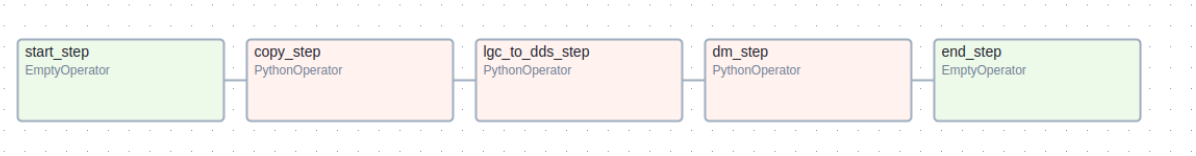
Вэб-интерфейс Airflow: http://localhost:8080
User: airflow

К контейнеру примонтирована папка для хранения ДАГов и исполняемых скриптов /home/mike/airflow/dags/:

- Код дага хранится в /home/mike/airflow/dags/big_dag-new.py
- Исполняемые скрипты:
  - /home/mike/airflow/dags/scripts/source_to_ods.py
  - /home/mike/airflow/dags/scripts/lgc_dds_with_cleaning.py

- /home/mike/airflow/dags/scripts/DM_tables.py
- /home/mike/airflow/dags/scripts/bad_data_collection.py

Описание ETL-потока:



ETL-поток реализован ДАГом Airflow **big_dag-new.py,** состоящим из трех тасок:

- **copy_step** - запускает скрипт **source_to_ods.py,** который с помощью библиотеки `psycopg` копирует исходные данные из слоя `source` исходной базы `SOURCE` в рабочую базу `etl_db_7`, слой `ods` для дальнейшей обработки.

  Между слоем `ods` и `dds` реализован логический слой `lgc` (в формате представлений) на котором названия таблиц и полей исходного слоя переименовываются на латиницу для удобства дальнейшей обработки.

- **lgc_to_dds_step** - запускает скрипт **lgc_dds_with_cleaning.py**, который с помощью библиотеки `psycopg` и набора `SQL-скриптов` забирает данные из слоя `lgc` и производит необходимую очистку и подготовку данных и сохраняет их в слой `dds`.

На этом этапе также происходит сбор “плохих” данных и перенос их в слой bad_data с добавлением причины забраковки. Скрипт: **bad_data_collection.py**

- **pre_dm**

- **dm_step** - запускает скрипт **DM-tables.py**, который с помощью библиотеки `psycopg` и набора `SQL-скриптов` забирает данные из слоя `dds`, производит необходимые вычисления и обработки и формирует две витрины в слое dm: ‘personal_data’ и ‘employee-skill’.

Описание хранилища данных

Хранилище данных реализовано в СУБД PostgreSQL со следующей структурой:

Слой	Схема	База данных	Параметры подключения к БД	Имя учетной записи
Source layer	source_data	source	host: 10.82.04 port: 5432	etl_user_7
ODS	ods	etl_db_7		
логический слой	lgc			
dds	dds			
bad_data	bad_data			
pre_dm	pre_dm			
data mart	dm			

Описание построения итоговой отчетности

Итоговая отчетность реализована с помощью Luxms BI с дашбордами, построенными на основе витрин данных из слоя dm/g_dm (data_mart)

3. Описание используемых сущностей с данными

Слой	Сущность
Source	базы_данных базы_данных_и_уровень_знаний_сотру инструменты инструменты_и_уровень_знаний_сотр образование_пользователей опыт_сотрудника_в_отраслях опыт_сотрудника_в_предметных_обла отрасли платформы платформы_и_уровень_знаний_сотруд предметная_область резюмедар сертификаты_пользователей сотрудники_дар среды_разработки среды_разработки_и_уровень_знаний_ технологии технологии_и_уровень_знаний_сотру

	типы_систем типы_систем_и_уровень_знаний_сотру уровень_образования уровни_владения_ин уровни_знаний уровни_знаний_в_отрасли уровни_знаний_в_предметной_област фреймворки фреймворки_и_уровень_знаний_сотру языки языки_пользователей языки_программирования языки_программирования_и_уровень
ods	базы_данных базы_данных_и_уровень_знаний_сотру инструменты инструменты_и_уровень_знаний_сотр образование_пользователей опыт_сотрудника_в_отраслях опыт_сотрудника_в_предметных_обла отрасли платформы платформы_и_уровень_знаний_сотруд предметная_область резюмедар сертификаты_пользователей сотрудники_дар среды_разработки среды_разработки_и_уровень_знаний_ технологии технологии_и_уровень_знаний_сотру типы_систем типы_систем_и_уровень_знаний_сотру уровень_образования уровни_владения_ин уровни_знаний уровни_знаний_в_отрасли уровни_знаний_в_предметной_област фреймворки фреймворки_и_уровень_знаний_сотру языки языки_пользователей языки_программирования языки_программирования_и_уровень
lgc	dbms_and_employee_grade dbms program_and_employee_grade program employee_education_level industry_employee_experience employee_domain_experience industry platform_and_employee_grade platform domain resume

	employee_certificate employee sde_and_employee_grade sde tool_and_employee_grade tool software_type_employee_grade software_type education_level foreign_language_level industry_experience experience grade framework_and_employee_grade framework employee_language programming_language_and_employee_gr ade programming_language language
dds	dbms_and_employee_grade dbms program_and_employee_grade program employee_education_level industry_employee_experience employee_domain_experience industry platform_and_employee_grade platform domain resume employee_certificate employee sde_and_employee_grade sde tool_and_employee_grade tool software_type_employee_grade software_type education_level foreign_language_level industry_experience experience grade framework_and_employee_grade framework employee_language programming_language_and_employee_gr ade programming_language language
bad_data	dbms_and_employee_grade dbms program_and_employee_grade program

	employee_education_level industry_employee_experience employee_domain_experience industry platform_and_employee_grade platform domain resume employee_certificate employee sde_and_employee_grade sde tool_and_employee_grade tool software_type_employee_grade software_type education_level foreign_language_level industry_experience experience grade framework_and_employee_grade framework employee_language programming_language_and_employee_grade programming_language language
g_pre_dm	employee employee_certificate employee_skill_grade employee_year_cer_flag grade skill
dm / g_dm	personal_data employee_skill skill grade employee employee_skill_grade