**Проектная работа по модулю**

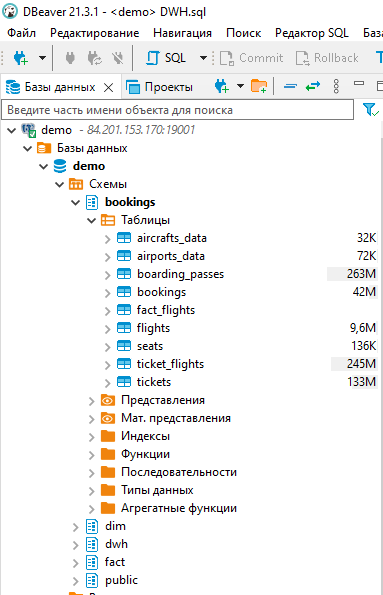
**“DWH”**

выполнила Дроздова Татьяна

**Описание ETL загрузки данных**

1 этап. Postgre SQL

Разворачиваем базу данных demo в Postgre SQL Создаем новую схему, в которой будет развернута модель данных.



Формируем таблицы дименшенов (измерений) для правильной нормализации данных и создание модели в формате «звезда»

Dim\_Calendar

Dim\_Passengers

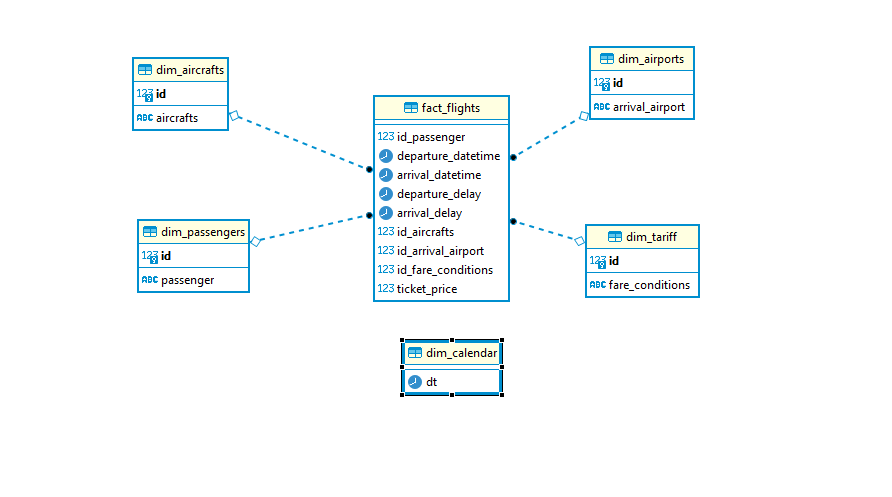
Dim\_Aircrafts

Dim\_Airports

Dim\_Tariff

Данные dim tables будут связаны с таблицей фактов на основании суррогатных ключей – id.

Создаем таблицу фактов Fact\_Flights. Поля итоговой таблицы заменяем на суррогатные id b связываем таблицы по ключам.



2 этап. Pentaho

Подключаем в рамках трансформации Pentaho к базе данных demo по принципу инъекции метаданных.

В рамках трансформации необходимо настроить выполнение проверки данных, например:

проверка формата данных (e.g.в поле цена не может лежать текст)

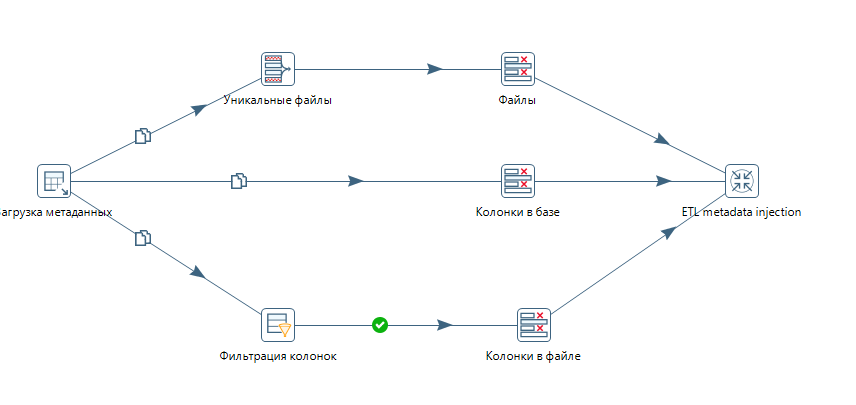
проверка диапазона дат (дата вылета меньше или равна дате прибытия)

проверка наличия дубликатов в dim tables

выбросы в значении цен на билеты в соответствии с категорией обслуживания

точечные проверки (аэропорт вылета и прилета не могут совпадать)

Некачественные строки записываются в отдельные лог-таблицы



3 этап. CSV

Данные после прохождения трансформации записываются в CSV-файлы. Идеальной была бы трансформация, которая записывала бы данные инкрементально, например, изменяя данные только за какой-то определенный промежуток времени.