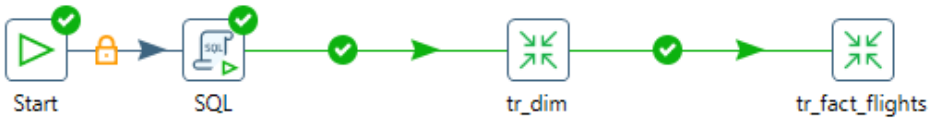
Общая структура ETL процедуры

Есть две основные трансформации - по загрузке данных в таблицы измерений и по загрузке данных в таблицу фактов fact-flights. Внутри них также есть проверки на качество данных по трем таблица (dim_passenger, dim_aircrafts, fact_flights), ошибочные строки собираются в отдельных rejected табилцах.

Их запускает отдельное задание, которые предварительно очищает все таблицы от данных.



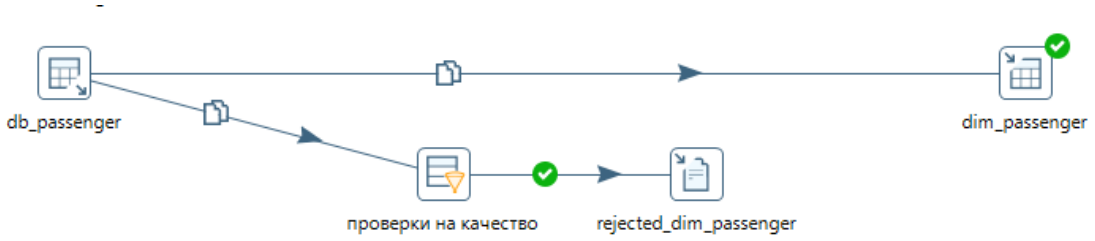
Трансформация по загрузке данных в таблицы измерений

Таблица измерений dim_date

Данные по справочнику дат генерируются через sql скрипт, в самой бд (файл в репозитории -)

id	date	ansi_date	day	week_number	month	year	week_day	holiday
20 100 101	2010-01-01	2010-01-01	5	53	1	2 009	1	0
20 100 102	2010-01-02	2010-01-02	6	53	1	2 009	0	0
20 100 103	2010-01-03	2010-01-03	7	53	1	2 009	0	0
20 100 104	2010-01-04	2010-01-04	1	1	1	2 010	1	0
20 100 105	2010-01-05	2010-01-05	2	1	1	2 010	1	0
20 100 106	2010-01-06	2010-01-06	3	1	1	2 010	1	0
20 100 107	2010-01-07	2010-01-07	4	1	1	2 010	1	0
20 100 108	2010-01-08	2010-01-08	5	1	1	2 010	1	0
20 100 109	2010-01-09	2010-01-09	6	1	1	2 010	0	0
20 100 110	2010-01-10	2010-01-10	7	1	1	2 010	0	0
20 100 111	2010-01-11	2010-01-11	1	2	1	2 010	1	0
20 100 112	2010-01-12	2010-01-12	2	2	1	2 010	1	0
20 100 113	2010-01-13	2010-01-13	3	2	1	2 010	1	0
20 100 114	2010-01-14	2010-01-14	4	2	1	2 010	1	0
20 100 115	2010-01-15	2010-01-15	5	2	1	2 010	1	0
20 100 116	2010-01-16	2010-01-16	6	2	1	2 010	0	0
20 100 117	2010-01-17	2010-01-17	7	2	1	2 010	0	0
20 100 118	2010-01-18	2010-01-18	1	3	1	2 010	1	0
20 100 119	2010-01-19	2010-01-19	2	3	1	2 010	1	0
20 100 120	2010-01-20	2010-01-20	3	3	1	2 010	1	0
20 100 121	2010-01-21	2010-01-21	4	3	1	2 010	1	0
20 100 122	2010-01-22	2010-01-22	5	3	1	2 010	1	0
20 100 123	2010-01-23	2010-01-23	6	3	1	2 010	0	0
20 100 124	2010-01-24	2010-01-24	7	3	1	2 010	0	0
20 100 125	2010-01-25	2010-01-25	1	4	1	2 010	1	0
20 100 126	2010-01-26	2010-01-26	2	4	1	2 010	1	0
20 100 127	2010-01-27	2010-01-27	3	4	1	2 010	1	0
20 100 128	2010-01-28	2010-01-28	4	4	1	2 010	1	0
20 100 129	2010-01-29	2010-01-29	5	4	1	2 010	1	0
20 100 130	2010-01-30	2010-01-30	6	4	1	2 010	0	0
20 100 131	2010-01-31	2010-01-31	7	4	1	2 010	0	0
20 100 201	2010-02-01	2010-02-01	1	5	2	2 010	1	0

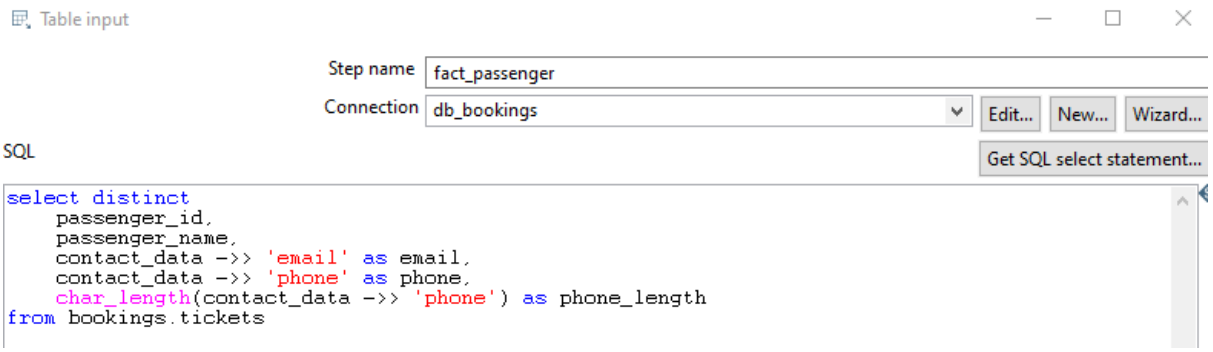
Таблица измерений dim_passenger



источник данных + процесс загрузки

Источник — бд demo, схема bookings, таблица tickets

Т.к. изначально в источнике контактные данные в формате jsonb, предварительно выцепляем телефон и email в отдельные столбцы



проверки на качество данных

Критерий качества данных	Описание проверки
Валидность	Если есть email, он должен содержать символ “@”
Валидность	Если есть телефон, он должен начинаться с символа “+”
Полнота	Кол-во символов в атрибуте phone - не должно превышать 12 символов (с учетом знака “+”)

Таблица измерений dim_airports

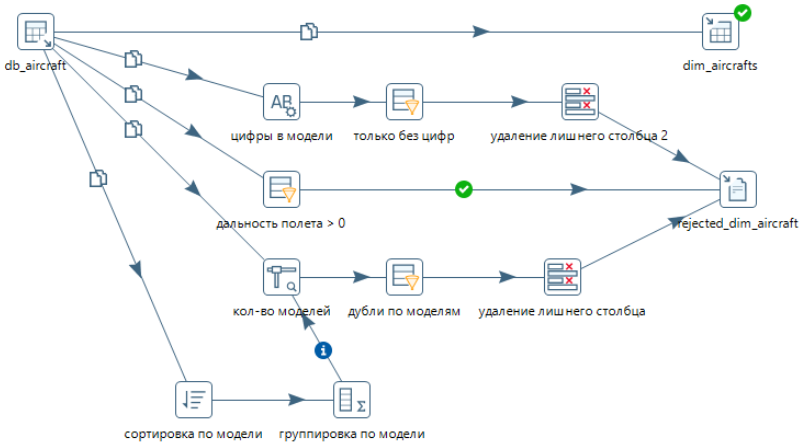
источник данных + процесс загрузки

Источник — бд demo, схема bookings, таблица airports_data

Т.к. изначально в источнике город и название аэропорта в формате jsonb, предварительно выцепляем только русские наименования. Также в отдельные столбцы разделяются широта и долгота.

```
select
    airport_code,
    airport_name ->>'ru'::text as airport_name,
    city ->>'ru'::text as city,
    coordinates[0] as longitude,
    coordinates[1] as latitude,
    timezone
from bookings.airports_data
```

Таблица измерений dim_aircrafts



источник данных + процесс загрузки

Источник — бд demo, схема bookings, таблица aircrafts_data.

Т.к. изначально в источнике название модели в формате jsonb, предварительно выцепляем только русские наименования.

```
select aircraft_code, model ->> 'ru' as model, range
from bookings.aircrafts_data
```

проверки на качество данных

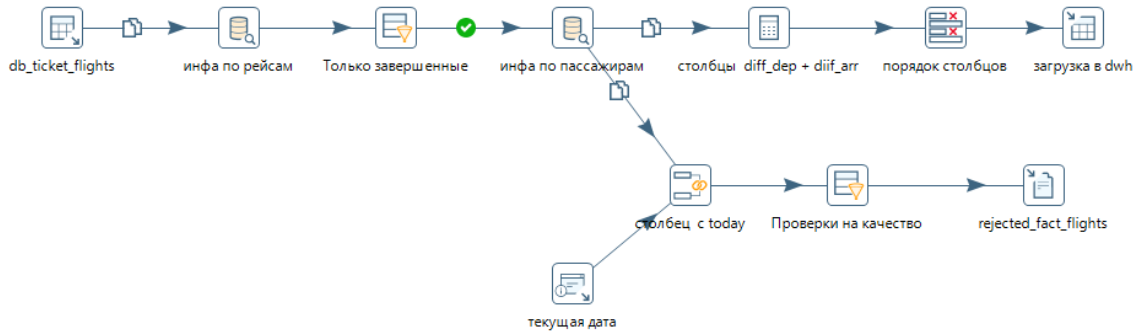
Критерий качества данных	Описание проверки
Валидность	Каждая модель должна содержать цифры в названии
Уникальность	Не должно быть повторяющихся названий модели
Валидность	Дальность полета д.б. больше нуля

Таблица измерений dim_tariff

источник данных

Источник — бд demo, схема bookings, таблица ticket_flights.

Трансформация по загрузке данных в таблицу фактов fact-flights



1) Шаг “db_ticket_flights”

Получаем данные из таблицы ticket_flights

2) Обогащаем данные из пред. шага след. информацией
    a) status
    b) scheduled_departure
    c) scheduled_arrival
    d) aircraft_code
    e) departure_airport
    f) arrival_airport
    g) actual_departure
    h) actual_arrival
3) Фильтруем только завершенные рейсы
4) Добавляем passenger_id на основе ticket_no
5) С помощью калькулятора добавляеются столбцы столбцы diff_dep + diif_arr (разница между фактической и запланированной датой вылета и прилета) в секундах
6) Определяем порядок столбцов
7) загружаем данные в таблицу fact_flights

проверки на качество данных

Критерий качества данных	Описание проверки
Достоверность	Дата фактического вылета не может быть больше сегодняшней даты
Достоверность	Аэропорт вылета и прилета не могут совпадать
Достоверность	Дата фактического вылета не может быть больше или равна даты фактического прилета