



**Проектная работа
“Реализация процедуры ETL”
по модулю
“ Data Warehouse”.**

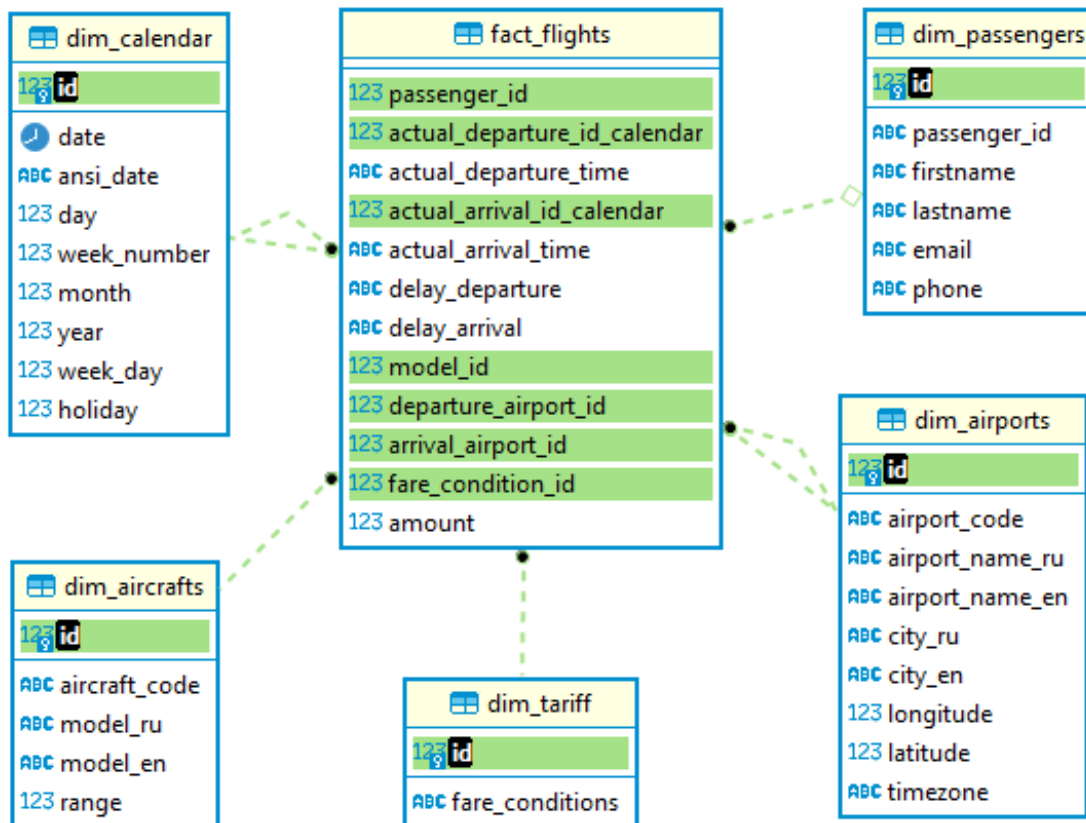
Оглавление.

ER-диаграмма – dim и fact таблицы в схеме dwh_bookings.....	3
Описание таблиц измерений	4
dim_passengers.....	4
dim_aircrafts	4
dim_airports.....	4
dim_calendar.....	4
dim_tariff	4
Описание таблицы фактов.....	5
fact_flights.....	5
ER-диаграмма – reject-таблицы в схеме dwh_bookings_reject	6
Описание reject-таблиц.....	7
reject_dim_passengers.....	7
reject_dim_aircrafts	7
reject_dim_tariff	7
reject_dim_airports.....	7
reject_fact_flights.....	8
Описание процедуры ETL	9
Трансформация dim_aircrafts.....	9
Трансформация dim_airoports	11
Трансформация dim_tariff.....	13
Трансформация dim_passengers.....	14
Трансформация fact_flights	17

Для выполнения проекта была создана база данных **dwh_demo_nj** и схемы:

- **bookings** – база данных с авиаперевозками по России. Содержит временной «срез» данных так, как будто в некоторый момент была сделана резервная копия реальной системы (Полное описание можно найти на странице <https://edu.postgrespro.ru/bookings.pdf>);
- **dwh_bookings** – хранилище данных с таблицами измерений и таблицей фактов;
- **dwh_bookings_reject** – reject _таблицы с данными, которые не прошли проверку качества.

ER-диаграмма – dim и fact таблицы в схеме dwh_bookings:



Таблицы измерений:

- dim_aircrafts - справочник самолетов;
- dim_airports - справочник аэропортов;
- dim_passengers - справочник пассажиров;
- dim_tariff - справочник тарифов (эконом/бизнес и тд);
- dim_calendar - справочник дат.

Таблица фактов:

- fact_flights - содержит совершенные перелеты, если в рамках билета был сложный маршрут с пересадками - каждый сегмент учитывается независимо.

Описание таблиц измерений.

dim_passengers. Источник данных - таблица tickets.

id serial4 primary key	Суррогатный идентификатор.
passenger_id varchar(20) unique not null	Идентификатор пассажира.
firstname varchar(70) not null	Имя пассажира.
lastname varchar(70) not null	Фамилия пассажира.
email varchar(50) not null	Email-адрес.
phone varchar(50) not null	Номер телефона.

dim_aircrafts. Источник данных - таблица aircrafts_data.

id serial4 primary key	Суррогатный идентификатор.
aircraft_code bpchar(3) not null unique	Трёхзначный код самолета.
model_ru varchar(300) not null	Модель самолета на русском языке.
model_en varchar(300) not null	Модель самолета на английском языке.
range int4 not null	Максимальная дальность полёта (км).

dim_airports. Источник данных - таблица airports_data.

id serial4 primary key	Суррогатный идентификатор.
airport_code bpchar(3) not null unique	Трёхзначный код аэропорта.
airport_name_ru varchar(200) not null	Название аэропорта на русском языке.
airport_name_en varchar(200) not null	Название аэропорта на английском языке.
city_ru varchar(100) not null	Название города на русском языке.
city_en varchar(100) not null	Название города на английском языке.
longitude decimal not null	Географическая долгота в градусах.
latitude decimal not null	Географическая широта в градусах.
timezone text not null	Временная зона аэропорта.

dim_calendar. Генерируется с помощью запроса SQL.

id int4 primary key	Суррогатный идентификатор в формате 'yyyymmdd'.
date date	Дата в формате 'yyyy-mm-dd'.
day int4	Номер дня недели.
week_number int4	Номер недели.
month int4	Номер месяца.
year int4	Год.
week_day int4	Выходной или нет (0 или 1).
holiday int4	Праздник или нет (0 или 1).

dim_tariff. Источник данных - таблица seats.

id serial4 primary key	Суррогатный идентификатор.
fare_conditions varchar(10) not null unique	Класс обслуживания.

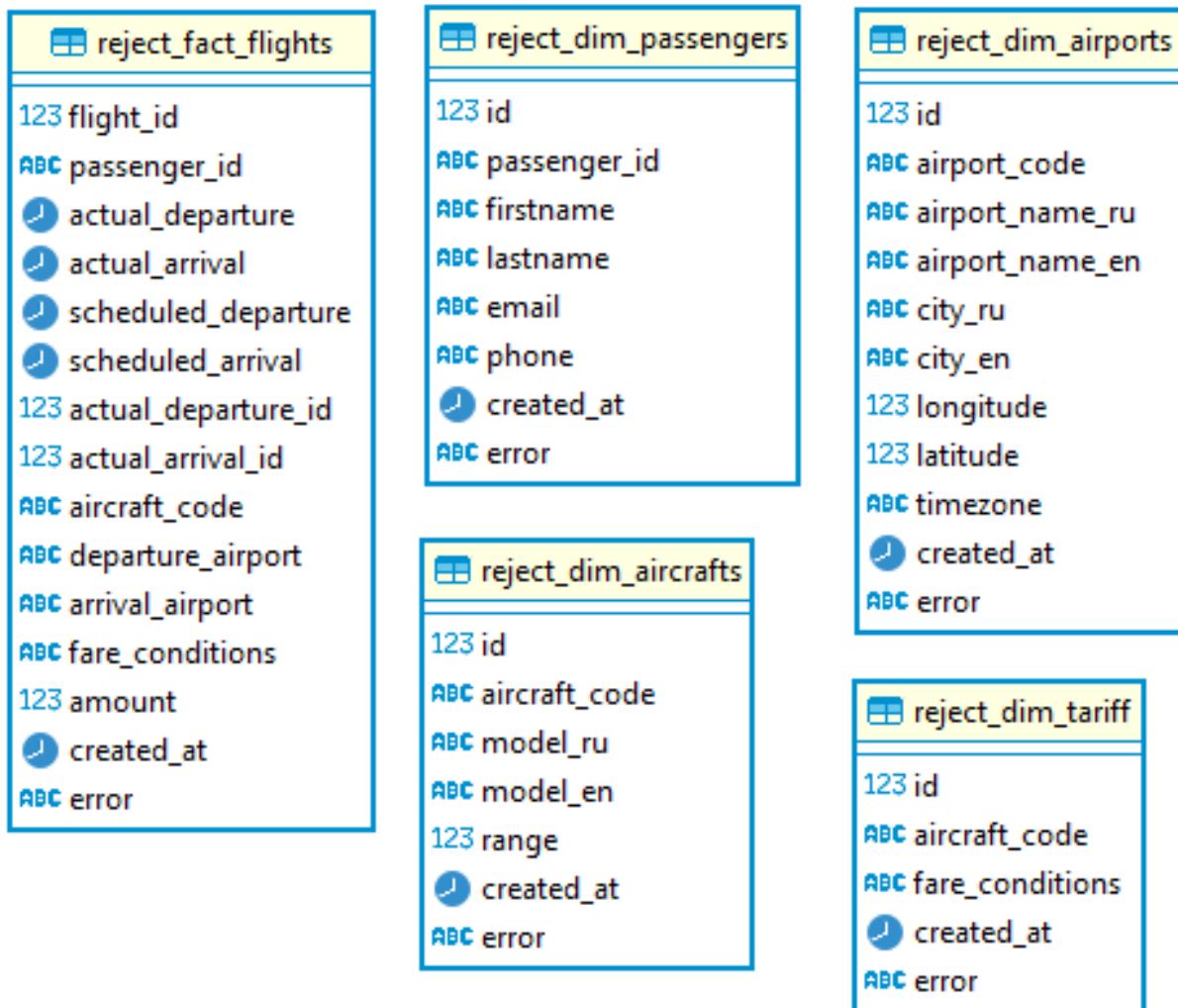
Описание таблицы фактов.

fact_flights. Источники данных – таблицы flights, ticket_flights, tickets.

passenger_id int not null	Суррогатный идентификатор пассажира. Связь с таблицей dim_passengers(id).
actual_departure_id_calendar int not null	Суррогатный идентификатор даты фактического вылета. Связь с таблицей dim_calendar(id).
actual_departure_time varchar(20) not null	Фактические дата и время вылета.
actual_arrival_id_calendar int not null	Суррогатный идентификатор даты фактического прилета. Связь с таблицей dim_calendar(id).
actual_arrival_time varchar(20) not null	Фактические дата и время прилета.
delay_departure varchar(20) not null	Задержка вылета (разница между фактической и запланированной датой в секундах).
delay_arrival varchar(20) not null	Задержка прилета (разница между фактической и запланированной датой в секундах).
model_id int not null	Суррогатный идентификатор модели самолета. Связь с таблицей dim_aircrafts(id).
departure_airport_id int not null	Суррогатный идентификатор аэропорта отправления. Связь с таблицей dim_airports(id).
arrival_airport_id int not null	Суррогатный идентификатор аэропорта прибытия. Связь с таблицей dim_airports(id).
fare_condition_id int not null	Суррогатный идентификатор класса обслуживания. Связь с таблицей dim_tariff(id).
amount numeric(10,2) not null	Стоимость перелета.

Строки, которые не прошли проверку качества, записываются в отдельные reject-таблицы.

ER-диаграмма – reject-таблицы в схеме dwh_bookings_reject :



Reject-таблицы:

- reject_dim_passengers - данные из таблицы dim_passengers.
- reject_dim_aircrafts - данные из таблицы dim_aircrafts.
- reject_dim_airports данные из таблицы dim_airports.
- reject_dim_tariff - данные из таблицы dim_tariff.
- reject_fact_flights - данные из таблицы fact_flights.

Описание reject-таблиц.

reject_dim_passengers.

id serial4 primary key	Суррогатный идентификатор.
passenger_id varchar(20)	Идентификатор пассажира.
firstname varchar(70)	Имя пассажира.
lastname varchar(70)	Фамилия пассажира.
email varchar(50)	Email-адрес.
phone varchar(50)	Номер телефона.
created_at timestamp	Дата и время регистрации ошибки.
error text	Описание ошибки валидации.

reject_dim_aircrafts.

id serial4	Суррогатный идентификатор.
aircraft_code bpchar(3)	Трёхзначный код самолета.
model_ru	Модель самолета на русском языке.
model_en varchar(300)	Модель самолета на английском языке.
range int4	Максимальная дальность полёта (км).
created_at timestamp	Дата и время регистрации ошибки.
error text	Описание ошибки валидации.

reject_dim_tariff.

id serial4	Суррогатный идентификатор.
fare_conditions varchar(10)	Класс обслуживания.
aircraft_code bpchar(3)	Трёхзначный код самолета.
created_at timestamp	Дата и время регистрации ошибки.
error text	Описание ошибки валидации.

reject_dim_airports.

id serial4	Суррогатный идентификатор.
airport_code bpchar(3)	Трёхзначный код аэропорта.
airport_name_ru varchar(200)	Название аэропорта на русском языке.
airport_name_en varchar(200)	Название аэропорта на английском языке.
city_ru varchar(100)	Название города на русском языке.
city_en varchar(100)	Название города на английском языке.
longitude decimal	Географическая долгота в градусах.
latitude decimal	Географическая широта в градусах.
timezone text	Временная зона аэропорта.
created_at timestamp	Дата и время регистрации ошибки.
error text	Описание ошибки валидации.

reject_fact_flights.

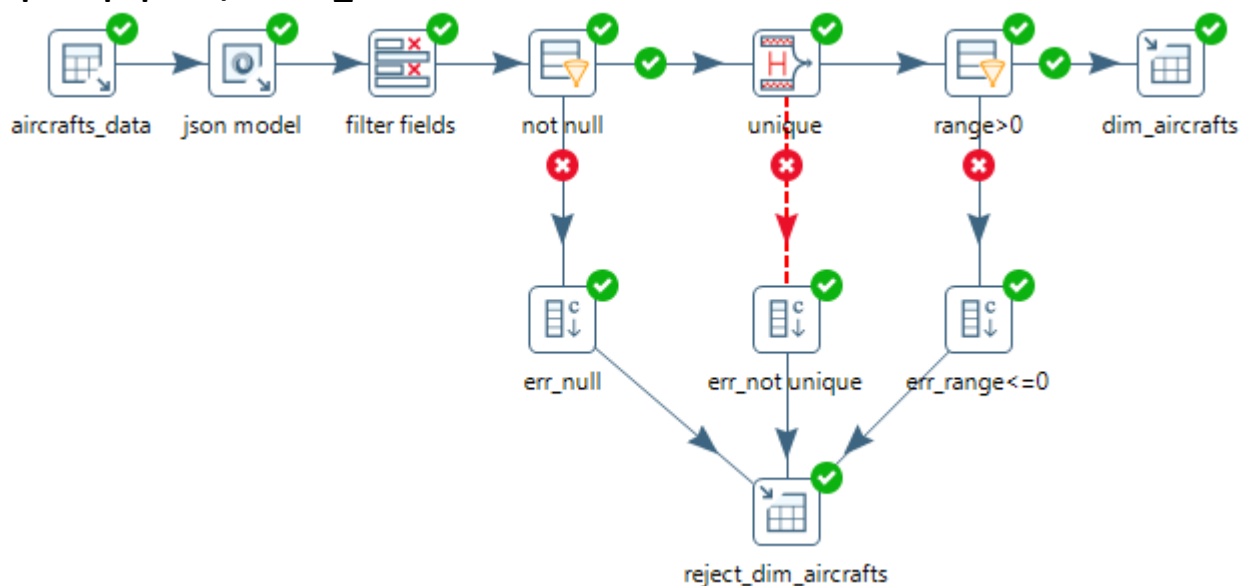
flight_id int	Идентификатор рейса.
passenger_id varchar(20)	Идентификатор пассажира.
actual_departure timestampz	Фактические дата и время вылета.
actual_arrival timestampz	Фактические дата и время прилета.
scheduled_departure timestampz	Время вылета по расписанию.
scheduled_arrival timestampz	Время прилета по расписанию.
actual_departure_id int	Идентификатор фактической даты вылета в формате 'YYYYMMDD'.
actual_arrival_id int	Идентификатор фактической даты прилета в формате 'YYYYMMDD'.
aircraft_code bpchar(3)	Трехзначный код самолета.
departure_airport bpchar(3)	Трехзначный код аэропорта отправления.
arrival_airport bpchar(3)	Трехзначный код аэропорта прибытия.
fare_conditions varchar(10)	Класс обслуживания.
amount numeric(10, 2)	Стоимость перелета.
created_at timestamp	Дата и время регистрации ошибки.
error text	Описание ошибки валидации.






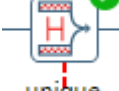
Описание процедуры ETL.


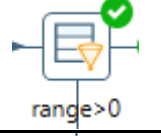
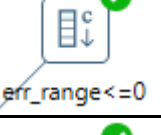
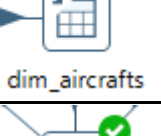
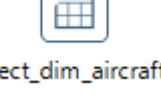
Процесс загрузки данных в хранилище состоит из выполнения 5 трансформаций (dim_aircrafts.ktr, dim_airoports.ktr, dim_passengers.ktr, dim_tariff.ktr, fact_flights.ktr) объединенных в задании start_bookings_FW_DWH_2023.kjb.



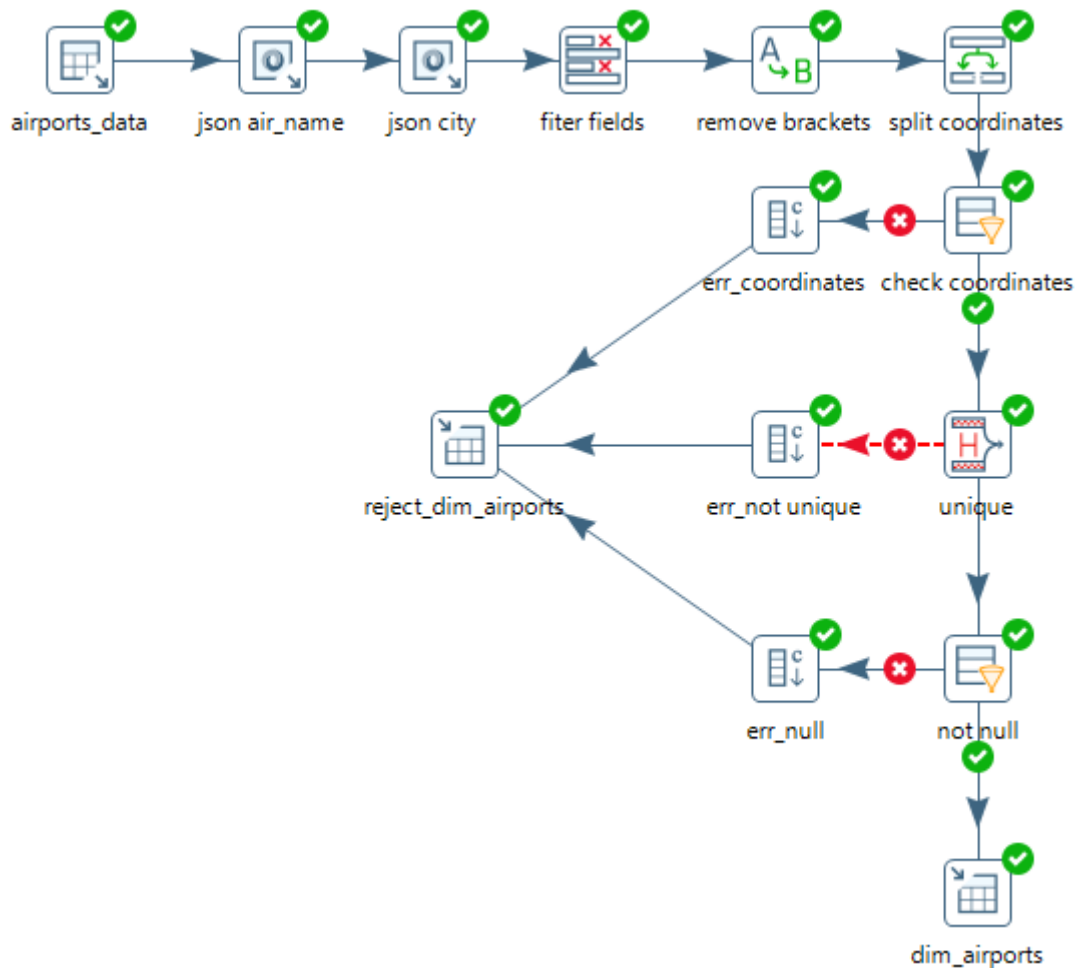
Трансформация dim_aircrafts.









Название шага трансформации	Тип	Описание
 aircrafts_data	Table input	Получение данных из таблицы aircrafts_data и текущей даты.
 json model	Json input	Извлечение названия самолета на английском и русском языках из JSON.
 filter fields	Select values	Удаление лишних полей с данными.
 not null	Filter rows	Фильтрация строк по наличию пустых данных.
 err_null	Add constants	Добавление поля "error" с указанием на наличие пустых данных.
 unique	Unique rows (HashSet)	Проверка на уникальность данных в поле "aircraft_code".

 err_notunique	Add constants	Добавление поля "error" с указанием на наличие неуникальных данных.
 range>0	Filter rows	Фильтрация данных по полю "range". Максимальная дальность полета должна быть больше нуля.
 err_range<=0	Add constants	Добавление поля "error" с указанием, что "range" <= 0.
 dim_aircrafts	Table output	Запись данных в таблицу измерений dim_aircrafts.
 reject_dim_aircrafts	Table output	Запись некачественных данных в таблицу reject_dim_aircrafts.

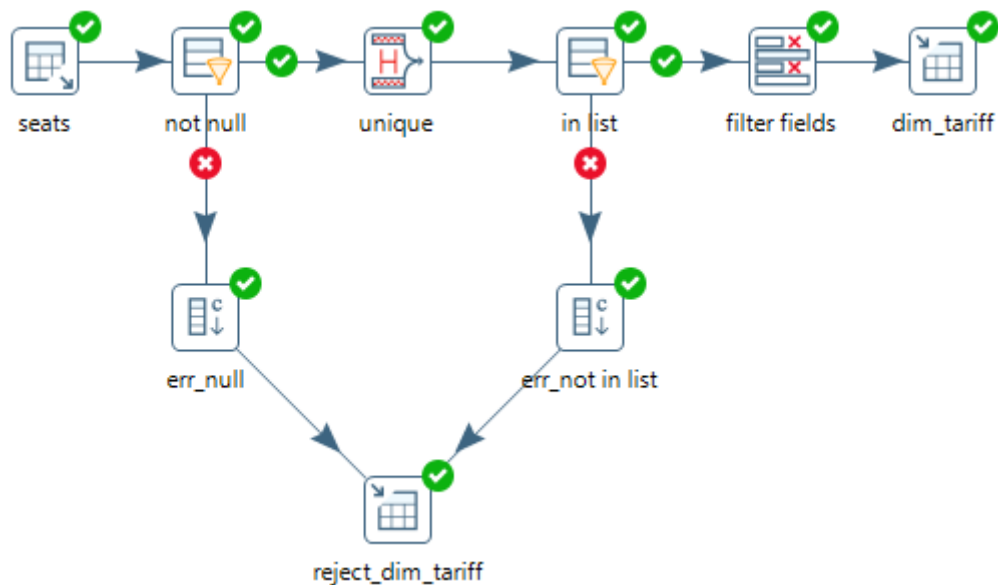
Трансформация dim_airports.



Название шага трансформации	Тип	Описание
 airports_data	Table input	Получение данных из таблицы aircrafts_data и текущей даты.
 json air_name	Json input	Извлечение названия аэропорта на русском и английском языках из JSON.
 json city	Json input	Извлечение названия городов на русском и английском языках из JSON.
 fiter fields	Select values	Удаление лишних полей с данными.
 remove brackets	Replace in string	Удаление круглых скобок в поле с координатами "coordinates".
 split coordinates	Split fields	Разделение координат на два поля – "longitude" (долгота) и "latitude" (широта).

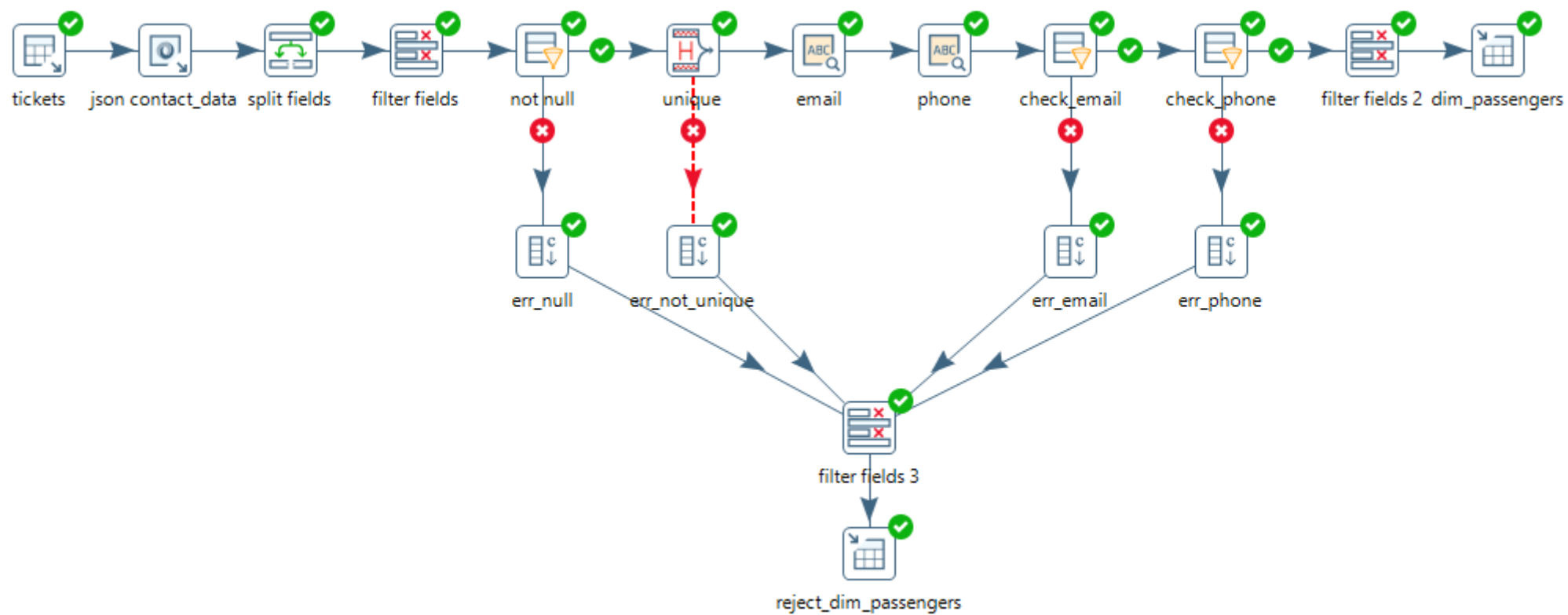
 check coordinates	Filter rows	Фильтрация координат по их корректности (-180<="longitude"<=180 и -90<="latitude"<=90)
 err_coordinates	Add constants	Добавление поля "error" с указанием на некорректность координат.
 unique	Unique rows (HashSet)	Проверка на уникальность данных в поле "airport_code".
 err_not unique	Add constants	Добавление поля "error" с указанием на наличие неуникальных данных.
 not null	Filter rows	Фильтрация строк по наличию пустых данных.
 err_null	Add constants	Добавление поля "error" с указанием на наличие пустых данных.
 reject_dim_airports,	Table output	Запись некачественных данных в таблицу reject_dim_airports.
 dim_airports	Table output	Запись данных в таблицу измерений dim_airports.

Трансформация dim_tariff.






Название шага трансформации	Тип	Описание
seats	Table input	Получение данных из таблицы seats и текущей даты.
not null	Filter rows	Фильтрация строк по наличию пустых данных.
err_null	Add constants	Добавление поля “error” с указанием на наличие пустых данных.
unique	Unique rows (HashSet)	Проверка на уникальность данных в поле “fare_conditions”.
in list	Filter rows	Фильтрация данных по соответствию указанного класса обслуживания со списком (Economy, Comfort, Business).
err_not in list	Add constants	Добавление поля “error” с указанием на несоответствие требуемому классу обслуживания.
reject_dim_tariff	Table output	Запись некачественных данных в таблицу reject_dim_tariff.
filter fields	Select values	Удаление лишних полей с данными.
dim_tariff	Table output	Запись данных в таблицу измерений dim_tariff.

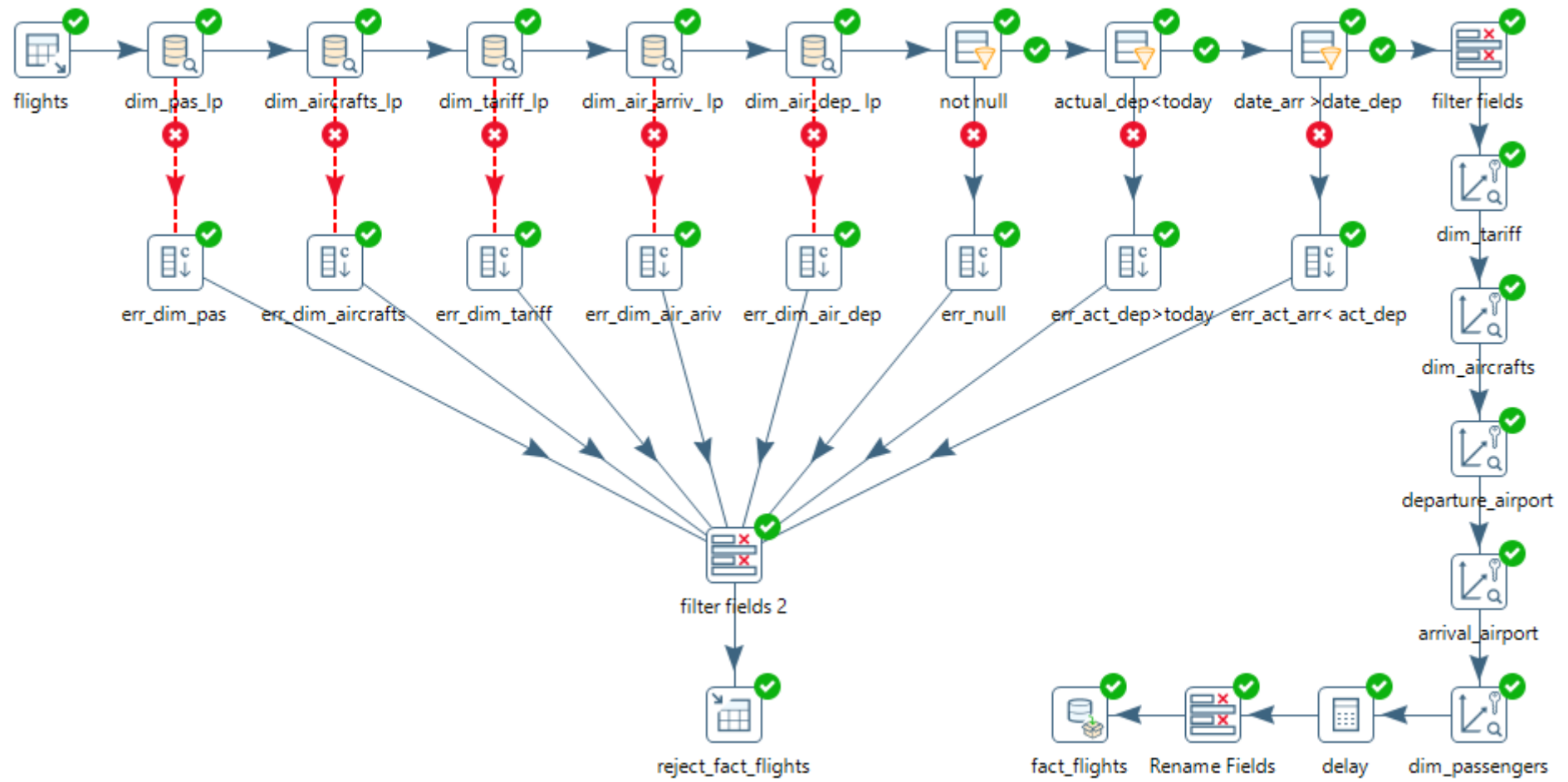
Трансформация dim_passengers.






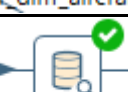

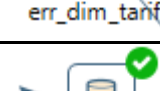

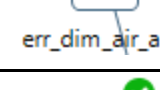

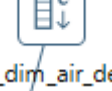








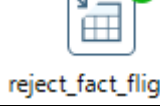

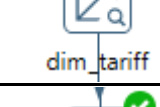

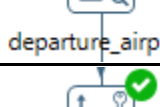

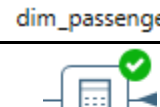
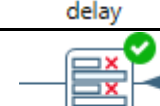


Название шага трансформации	Тип	Описание
 tickets	Table input	Получение данных из таблицы tickets и текущей даты.
 json contact_data	Json input	Извлечение "email" и "phone" из JSON.
 split fields	Split fields	Разделение данных в поле "passenger_name" на два поля "firstname" и "lastname".
 filter fields	Select values	Удаление лишних полей с данными.
 not null	Filter rows	Фильтрация строк по наличию пустых данных.
 err_null	Add constants	Добавление поля "error" с указанием на наличие пустых данных.
 unique	Unique rows (HashSet)	Проверка на уникальность данных в поле "passenger_id".
 err_not_unique	Add constants	Добавление поля "error" с указанием на наличие неуникальных данных.
 email	Regex evaluation	Проверка соответствия email-адреса с помощью регулярного выражения.
 phone	Regex evaluation	Проверка соответствия номера телефона с помощью регулярного выражения.
 check_email	Filter rows	Фильтрация строк с некорректным форматом email-адреса.
 err_email	Add constants	Добавление поля "error" с указанием на некорректность формата email-адреса.
 check_phone	Filter rows	Фильтрация строк с некорректным форматом номера телефона.
 err_phone	Add constants	Добавление поля "error" с указанием на некорректность формата номера телефона.

 filter fields 3	Select values	Удаление лишних полей с данными.
 reject_dim_passengers	Table output	Запись некачественных данных в таблицу reject_dim_passengers.
 filter fields 2	Select values	Удаление лишних полей с данными.
 dim_passengers	Table output	Запись данных в таблицу измерений dim_passengers.

Трансформация fact_flights.



Название шага трансформации	Тип	Описание
 flights	Table input	Получение данных из таблиц flights, ticket_flights, tickets и текущей даты.
 dim_pass_lp	Database lookup	Проверка соответствия данных таблицы flights с данными таблицы dim_passengers по ключу passenger_id.
 err_dim_pas	Add constants	Добавление поля "error" с указанием, что поля passenger_id таблицы flights не найдены в таблице dim_passengers.
 dim_aircrafts_lp	Database lookup	Проверка соответствия данных таблицы flights с данными таблицы dim_aircrafts по ключу aircraft_code.
 err_dim_aircrafts	Add constants	Добавление поля "error" с указанием, что поле aircraft_code таблицы flights не найдено в таблице dim_aircrafts.
 dim_tariff_lp	Database lookup	Проверка соответствия данных таблицы flights+ticket_flights с данными таблицы dim_tariff по ключу fare_conditions.
 err_dim_tariff	Add constants	Добавление поля "error" с указанием, что поле fare_conditions таблицы flights+ticket_flights не найдено в таблице dim_tariff.
 dim_air_arriv_lp	Database lookup	Проверка соответствия данных таблицы flights с данными таблицы dim_airports по ключам airport_code и arrival_airport.
 err_dim_air_ariv	Add constants	Добавление поля "error" с указанием, что значение поля arrival_airport таблицы flights не найдено в таблице dim_airports (airport_code).
 dim_air_dep_lp	Database lookup	Проверка соответствия данных поля departure_airport таблицы flights с данными поля airport_code таблицы dim_airports.
 err_dim_air_dep	Add constants	Добавление поля "error" с указанием, что значение поля departure_airport таблицы flights не найдено в таблице dim_airports (airport_code).
 not null	Filter rows	Фильтрация строк по наличию пустых данных.
 err_null	Add constants	Добавление поля "error" с указанием на наличие пустых данных.
 actual_dep < today	Filter rows	Фильтрация строк, в которых фактическая дата вылета меньше текущей даты.

 err_act_dep > today	Add constants	Добавление поля "error" с указанием, что фактическая дата вылета больше текущей даты.
 date_arr > date_dep	Filter rows	Фильтрация строк, в которых фактическая дата прилета больше фактической даты вылета.
 err_act_arr < act_dep	Add constants	Добавление поля "error" с указанием, что фактическая дата прилета меньше фактической даты вылета.
 filter fields 2	Select values	Удаление лишних полей с данными.
 reject_fact_flights	Table output	Запись некачественных данных в таблицу reject_fact_flights.
 filter fields	Select values	Удаление лишних полей с данными.
 dim_tariff	Combination lookup/update	Создание суррогатного ключа для поля fare_conditions.
 dim_aircrafts	Combination lookup/update	Создание суррогатного ключа для поля aircraft_code.
 departure_airport	Combination lookup/update	Создание суррогатного ключа для поля departure_airport.
 arrival_airport	Combination lookup/update	Создание суррогатного ключа для поля arrival_airport.
 dim_passengers	Combination lookup/update	Создание суррогатного ключа для поля passenger_id.
 delay	Calculator	Расчет задержки вылета и прилета (разница между фактической и запланированной датой в секундах).
 Rename Fields	Select values	Удаление лишних полей, переименование суррогатных ключей.
 fact_flights	PostgreSQL bulk loader	Запись данных в таблицу фактов fact_flights.