****

**Проектная работа**

**“Реализация процедуры ETL”**

**по модулю**

**“Data Warehouse”.**

**Оглавление.**

[**ER-диаграмма –** dim и fact таблицы в схеме dwh\_bookings 3](#_Toc130213893)

[**Описание таблиц измерений** 4](#_Toc130213894)

[dim\_passengers 4](#_Toc130213895)

[dim\_aircrafts 4](#_Toc130213896)

[dim\_airports. 4](#_Toc130213897)

[dim\_calendar 4](#_Toc130213898)

[dim\_tariff 4](#_Toc130213899)

[**Описание таблицы фактов** 5](#_Toc130213900)

[fact\_flights 5](#_Toc130213901)

[**ER-диаграмма –** reject-таблицы в схеме dwh\_bookings\_reject 6](#_Toc130213902)

[**Описание reject-таблиц** 7](#_Toc130213903)

[reject\_dim\_passengers 7](#_Toc130213904)

[reject\_dim\_aircrafts 7](#_Toc130213905)

[reject\_dim\_tariff 7](#_Toc130213906)

[reject\_dim\_airports 7](#_Toc130213907)

[reject\_fact\_flights 8](#_Toc130213908)

[**Описание процедуры ETL** 9](#_Toc130213909)

[Трансформация dim\_aircrafts 9](#_Toc130213910)

[Трансформация dim\_airoports 11](#_Toc130213911)

[Трансформация dim\_tariff 13](#_Toc130213912)

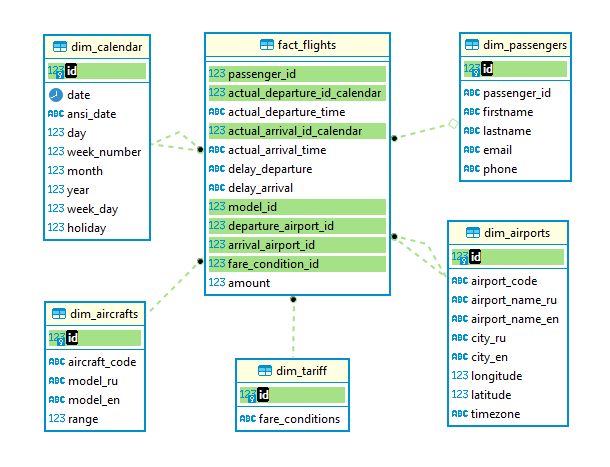
[Трансформация dim\_passengers 14](#_Toc130213913)

[Трансформация fact\_flights 17](#_Toc130213914)

Для выполнения проекта была создана база данных **dwh\_demo\_nj** и схемы:

* **bookings** – база данных с авиаперевозками по России. Cодержит 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. |

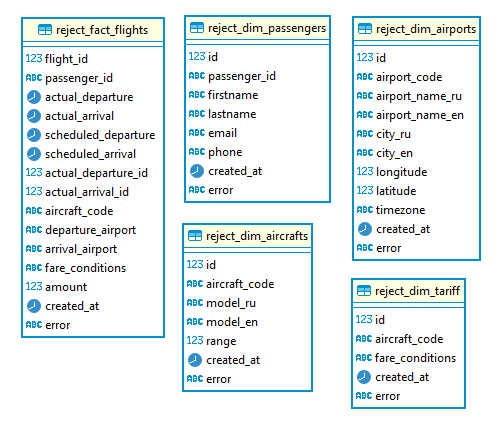
# **Описание таблицы фактов.**

## **fact\_flights.** Источники данных – таблицы flights, ticket\_flights, tickets.

|  |  |
| --- | --- |
| passenger\_id int not null | Суррогатный идентификатор пассажира.  Cвязь с таблицей dim\_passengers(id). |
| actual\_departure\_id\_calendar int not null | Суррогатный идентификатор даты фактического вылета.  Cвязь 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. |
| сreated\_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). |
| сreated\_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. |
| сreated\_at timestamp | Дата и время регистрации ошибки. |
| error text | Описание ошибки валидации. |

## **reject\_dim\_airports**.

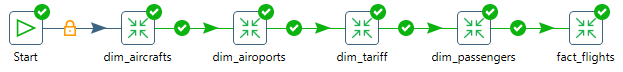
|  |  |
| --- | --- |
| id serial4 | Суррогатный идентификатор. |
| airport\_code bpchar(3) | Трёхзначный код аэропорта. |
| airport\_name\_ru varchar(200) | Название аэропорта на русском языке. |
| airport\_name\_en varchar(200) | Название аэропорта на английском языке. |
| city\_ru varchar(100) | Название 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. |
| сreated\_at timestamp | Дата и время регистрации ошибки. |
| error text | Описание ошибки валидации. |

## **reject\_fact\_flights**.

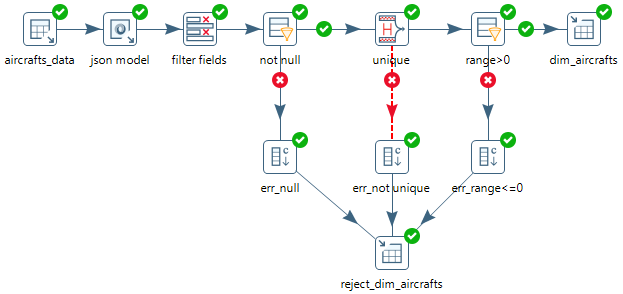
|  |  |
| --- | --- |
| flight\_id int | Идентификатор рейса. |
| passenger\_id varchar(20) | Идентификатор пассажира. |
| actual\_departure timestamptz | Фактические дата и время вылета. |
| actual\_arrival timestamptz | Фактические дата и время прилета. |
| scheduled\_departure timestamptz | Время вылета по расписанию. |
| scheduled\_arrival timestamptz | 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. |
| сreated\_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.

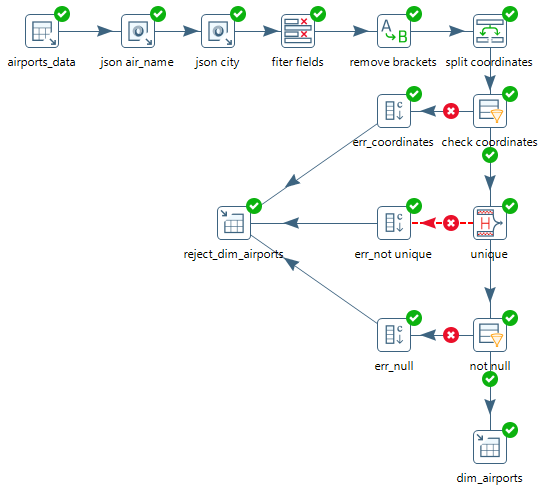
****

## **Трансформация dim\_aircrafts.**



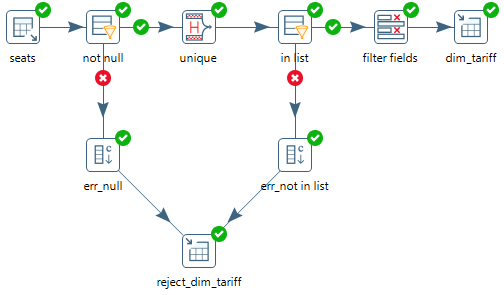
|  |  |  |
| --- | --- | --- |
| **Название шага трансформации** | **Тип** | **Описание** |
|  | Table input | Получение данных из таблицы aircrafts\_data и текущей даты. |
|  | Json input | Извлечение названия самолета на английском и русском языках из JSON. |
|  | Select values | Удаление лишних полей с данными. |
|  | Filter rows | Фильтрация строк по наличию пустых данных. |
|  | Add constants | Добавление поля “error” с указанием на наличие пустых данных. |
|  | Unique rows (HashSet) | Проверка на уникальность данных в поле “aircraft\_code”. |
|  | Add constants | Добавление поля “error” с указанием на наличие неуникальных данных. |
|  | Filter rows | Фильтрация данных по полю “range”. Максимальная дальность полета должна быть больше нуля. |
|  | Add constants | Добавление поля “error” с указанием, что  “range” <= 0. |
|  | Table output | Запись данных в таблицу измерений dim\_aircrafts. |
|  | Table output | Запись некачественных данных в таблицу reject\_dim\_aircrafts. |

## **Трансформация dim\_airoports.**



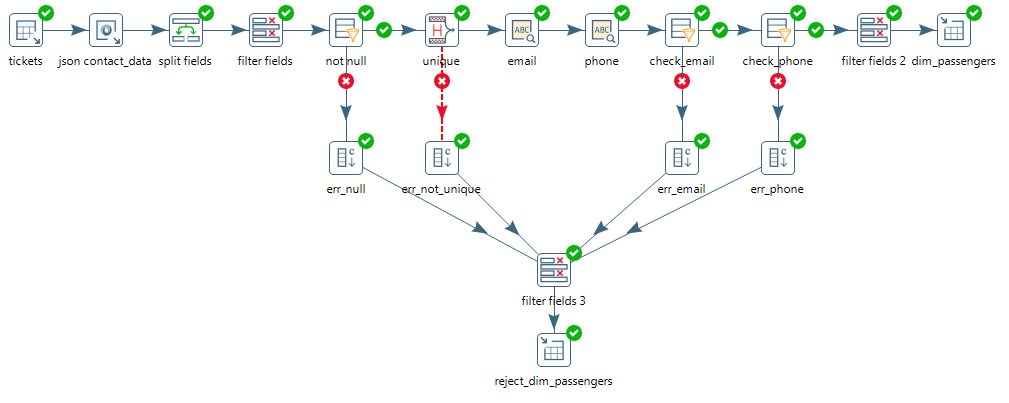
|  |  |  |
| --- | --- | --- |
| **Название шага трансформации** | **Тип** | **Описание** |
|  | Table input | Получение данных из таблицы aircrafts\_data и текущей даты. |
|  | Json input | Извлечение названия аэропорта на русском и английском языках из JSON. |
|  | Json input | Извлечение названия городов на русском и английском языках из JSON. |
|  | Select values | Удаление лишних полей с данными. |
|  | Replace in string | Удаление круглых скобок в поле с координатами “coordinates”. |
|  | Split fields | Разделение координат на два поля – “longitude” (долгота) и “latitude” (широта). |
|  | Filter rows | Фильтрация координат по их корректности  (-180<=“longitude”<=180 и -90<=“latitude”<=90) |
|  | Add constants | Добавление поля “error” с указанием на некорректность координат. |
|  | Unique rows (HashSet) | Проверка на уникальность данных в поле “airport\_code ”. |
|  | Add constants | Добавление поля “error” с указанием на наличие неуникальных данных. |
|  | Filter rows | Фильтрация строк по наличию пустых данных. |
|  | Add constants | Добавление поля “error” с указанием на наличие пустых данных. |
|  | Table output | Запись некачественных данных в таблицу reject\_ dim\_airoports. |
|  | Table output | Запись данных в таблицу измерений dim\_airoports. |

## **Трансформация dim\_tariff.**



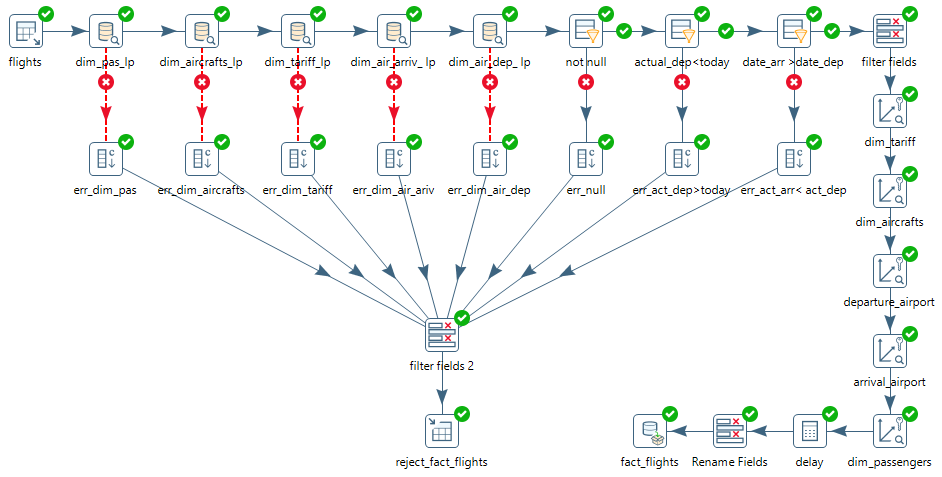
|  |  |  |
| --- | --- | --- |
| **Название шага трансформации** | **Тип** | **Описание** |
|  | Table input | Получение данных из таблицы seats и текущей даты. |
|  | Filter rows | Фильтрация строк по наличию пустых данных. |
|  | Add constants | Добавление поля “error” с указанием на наличие пустых данных. |
|  | Unique rows (HashSet) | Проверка на уникальность данных в поле “fare\_conditions ”. |
|  | Filter rows | Фильтрация данных по соответствию указанного класса обслуживания со списком (Economy, Comfort, Business). |
|  | Add constants | Добавление поля “error” с указанием на несоответствие требуемому классу обслуживания. |
|  | Table output | Запись некачественных данных в таблицу reject\_ dim\_tariff. |
|  | Select values | Удаление лишних полей с данными. |
|  | Table output | Запись данных в таблицу измерений dim\_tariff. |

**Трансформация dim\_passengers.**



|  |  |  |
| --- | --- | --- |
| **Название шага трансформации** | **Тип** | **Описание** |
|  | Table input | Получение данных из таблицы tickets и текущей даты. |
|  | Json input | Извлечение “email” и “phone” из JSON. |
|  | Split fields | Разделение данных в поле “passenger\_name” на два поля “firstname” и “lastname”. |
|  | Select values | Удаление лишних полей с данными. |
|  | Filter rows | Фильтрация строк по наличию пустых данных. |
|  | Add constants | Добавление поля “error” с указанием на наличие пустых данных. |
|  | Unique rows (HashSet) | Проверка на уникальность данных в поле “passenger\_id”. |
|  | Add constants | Добавление поля “error” с указанием на наличие неуникальных данных. |
|  | Regex evaluation | Проверка соответствия email-адреса с помощью регулярного выражения. |
|  | Regex evaluation | Проверка соответствия номера телефона с помощью регулярного выражения. |
|  | Filter rows | Фильтрация строк с некорректным форматом email-адреса. |
|  | Add constants | Добавление поля “error” с указанием на некорректность формата email-адреса. |
|  | Filter rows | Фильтрация строк с некорректным форматом номера телефона. |
|  | Add constants | Добавление поля “error” с указанием на некорректность формата номера телефона. |
|  | Select values | Удаление лишних полей с данными. |
|  | Table output | Запись некачественных данных в таблицу reject\_ dim\_passengers. |
|  | Select values | Удаление лишних полей с данными. |
|  | Table output | Запись данных в таблицу измерений dim\_passengers. |

## **Трансформация fact\_flights.**



|  |  |  |
| --- | --- | --- |
| **Название шага трансформации** | **Тип** | **Описание** |
|  | Table input | Получение данных из таблиц flights, ticket\_flights, tickets и текущей даты. |
|  | Database lookup | Проверка соответствия данных таблицы flights с данными таблицы dim\_passengers по ключу passenger\_id. |
|  | Add constants | Добавление поля “error” с указанием, что поля passenger\_id таблицы flights не найдены в таблице dim\_passengers. |
|  | Database lookup | Проверка соответствия данных таблицы flights с данными таблицы dim\_aircrafts по ключу aircraft\_code. |
|  | Add constants | Добавление поля “error” с указанием, что поле aircraft\_code таблицы flights не найдено в таблице dim\_aircrafts. |
|  | Database lookup | Проверка соответствия данных таблицы flights+ticket\_flights с данными таблицы dim\_tariff по ключу fare\_conditions. |
|  | Add constants | Добавление поля “error” с указанием, что поле fare\_conditions таблицы flights+ticket\_flights не найдено в таблице dim\_tariff. |
|  | Database lookup | Проверка соответствия данных таблицы flights с данными таблицы dim\_airports по ключам airport\_code и arrival\_airport. |
|  | Add constants | Добавление поля “error” с указанием, что значение поля arrival\_airport таблицы flights не найдено в таблице dim\_airports (airport\_code). |
|  | Database lookup | Проверка соответствия данных поля departure\_airport таблицы flights с данными поля airport\_code таблицы dim\_airports. |
|  | Add constants | Добавление поля “error” с указанием, что значение поля departure\_airport таблицы flights не найдено в таблице dim\_airports (airport\_code). |
|  | Filter rows | Фильтрация строк по наличию пустых данных. |
|  | Add constants | Добавление поля “error” с указанием на наличие пустых данных. |
|  | Filter rows | Фильтрация строк, в которых фактическая дата вылета меньше текущей даты. |
|  | Add constants | Добавление поля “error” с указанием, что фактическая дата вылета больше текущей даты. |
|  | Filter rows | Фильтрация строк, в которых фактическая дата прилета больше фактической даты вылета. |
|  | Add constants | Добавление поля “error” с указанием, что фактическая дата прилета меньше фактической даты вылета. |
|  | Select values | Удаление лишних полей с данными. |
|  | Table output | Запись некачественных данных в таблицу reject\_ fact\_flights. |
|  | Select values | Удаление лишних полей с данными. |
|  | Combination lookup/update | Создание суррогатного ключа для поля fare\_conditions. |
|  | Combination lookup/update | Создание суррогатного ключа для поля aircraft\_code. |
|  | Combination lookup/update | Создание суррогатного ключа для поля departure\_airport. |
|  | Combination lookup/update | Создание суррогатного ключа для поля arrival\_airport. |
|  | Combination lookup/update | Создание суррогатного ключа для поля passenger\_id. |
|  | Calculator | Расчет задержки вылета и прилета (разница между фактической и запланированной датой в секундах). |
|  | Select values | Удаление лишних полей, переименование суррогатных ключей. |
|  | PostgreSQL bulk loader | Запись данных в таблицу фактов fact\_flights. |