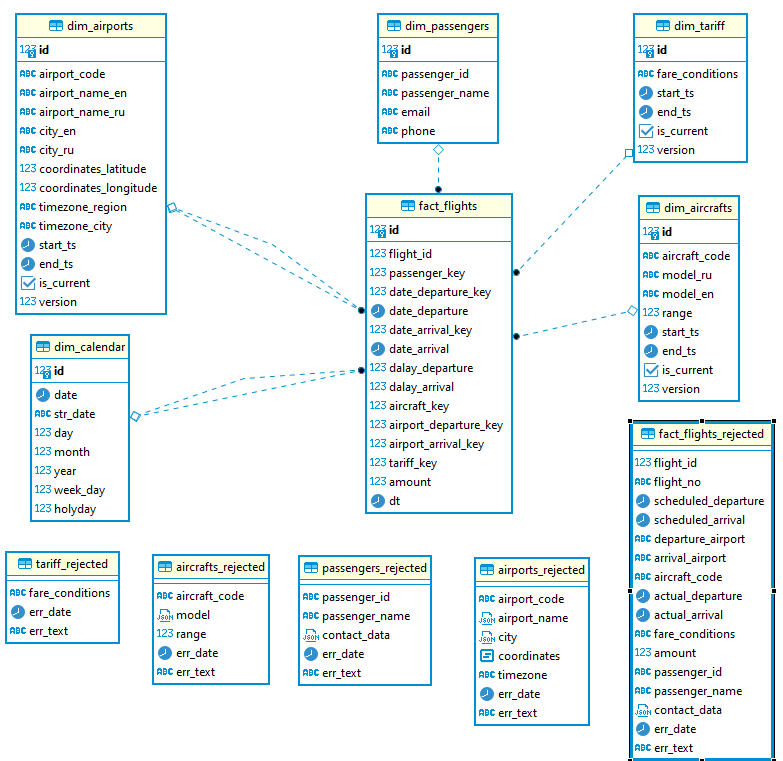
**Проектная работа по модулю “DWH”**

В рамках итоговой работы создана новая база данных со следующей структурой:



**Таблицы размерностей:**

dim\_calendar - справочник дат

dim\_passengers - справочник пассажиров

dim\_aircrafts - справочник самолетов

dim\_airports - справочник аэропортов

dim\_tariff - справочник тарифов (эконом/бизнес и тд)

**Таблица фактов**

fact\_flights - содержит совершенные перелеты, если в рамках билета был сложный маршрут с пересадками - каждый сегмент учитываем независимо

**Таблицы ошибок**

Содержат информацию по строкам, которые не прошли проверки качества данных и при обработке которых возникли ошибки, текст ошибки и дату загрузки данных.

dim\_calendar\_rejected - ошибки обработки dim\_calendar

dim\_passengers\_rejected - ошибки обработки dim\_passengers

dim\_aircrafts\_rejected - ошибки обработки dim\_aircrafts

dim\_airports\_rejected - ошибки обработки dim\_airports

dim\_tariff\_rejected - ошибки обработки dim\_tariff

fact\_flights\_rejected - ошибки обработки fact\_flights

**Описание ETL трансформаций**

**dim\_aircrafts**

Источник данных - таблица aircrafts\_data.

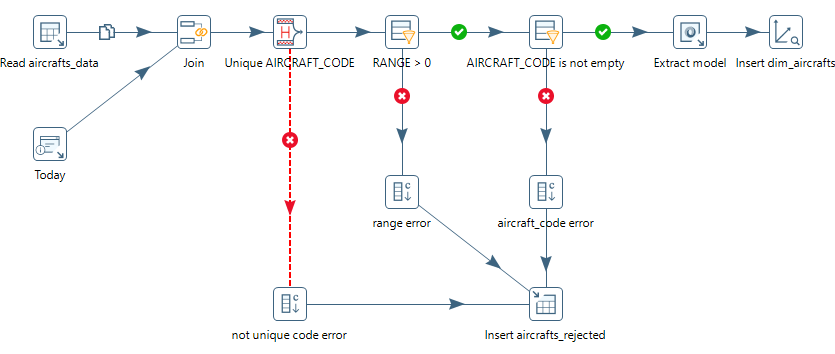
Для таблицы dim\_aircrafts реализованы следующие проверки качества данных:

1. значение поля range должно быть больше 0;

2. в поле aircraft\_code должны быть уникальные значения;

3. значения в поле aircraft\_code не могут быть пустыми.

Значения, не прошедшие проверки, записываются в таблицу aircrafts\_rejected.

****

| **Название шага трансформации** | **Тип** | **Описание** |
| --- | --- | --- |
|  | Table input | зачитка данных из таблицы источника bookings.aircrafts\_data |
|  | Get system info | получить текущую дату |
|  | JOIN rows | добавление Today в набор данных |
|  | Unique rowset (HashSet) | проверка значений поля aircraft\_code на уникальность |
|  | Add constants | для записей не прошедших проверку добавляем комментарий в поле err\_text о сути ошибки “Code is not unique” |
|  | Filter rows | проверка, что в поле range значение больше 0 |
|  | Add constants | для записей не прошедших проверку добавляем комментарий в поле err\_text о сути ошибки “Range is less or equal 0” |
|  | Filter rows | проверка, что поле aircraft\_code не пустое (NULL) и что его значение не равно пустой строке (“”) |
|  | Add constants | для записей не прошедших проверку добавляем комментарий в поле err\_text о сути ошибки “Code is empty” |
|  | Table output | запись в aircrafts\_rejected |
|  | JSON input | извлечение из данных поля model в model\_en и model\_ru |
|  | Dimension lookup/update | запись данных в хранилище, таблицу dim\_aircrafts |

**dim\_airports**

Источник данных - таблица airports\_data.

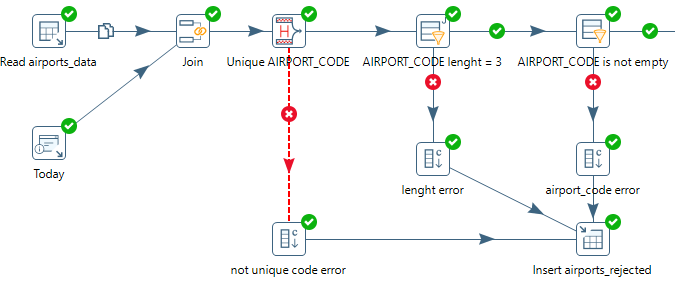
Для таблицы dim\_airports реализованы следующие проверки качества данных:

1. в поле airport\_code должны быть уникальные значения;

2. длина поля airport\_code должна быть равна 3 символам;

3. значения в поле airport\_code не могут быть пустыми.

Значения, не прошедшие проверки, записываются в таблицу airports\_rejected.



продолжение:



| **Название шага трансформации** | **Тип** | **Описание** |
| --- | --- | --- |
|  | Table input | зачитка данных из таблицы источника bookings.airports\_data |
|  | Get system info | получить текущую дату |
|  | JOIN rows | добавление Today в набор данных |
|  | Unique rowset (HashSet) | проверка значений поля airport\_code на уникальность |
|  | Add constants | для записей не прошедших проверку добавляем комментарий в поле err\_text о сути ошибки “Code is not unique” |
|  | Filter rows | проверка, что длина поля airport\_code равна 3 |
|  | Add constants | для записей не прошедших проверку добавляем комментарий в поле err\_text о сути ошибки “Lenght is not equal 3” |
|  | Filter rows | проверка, что поле airport\_code не пустое (NULL) и что его значение не равно пустой строке (“”) |
|  | Add constants | для записей не прошедших проверку добавляем комментарий в поле err\_text о сути ошибки “Code is empty” |
|  | Table output | запись в airports\_rejected |
|  | JSON input | извлечение из данных поля airport\_name в airport\_name\_en и airport\_name\_ru |
|  | JSON input | извлечение из данных поля city в city\_en и city\_ru |
|  | Split fields | разделение значения из поля timezone в поля timezone\_region и timezone\_city |
|  | Replace in string | удаление из поля coordinates скобок |
|  | Split fields | разделение значения из поля coordinates в поля coordinates\_latitude и coordinates\_longitude |
|  | Dimension lookup/update | запись данных в хранилище, таблицу dim\_airports |

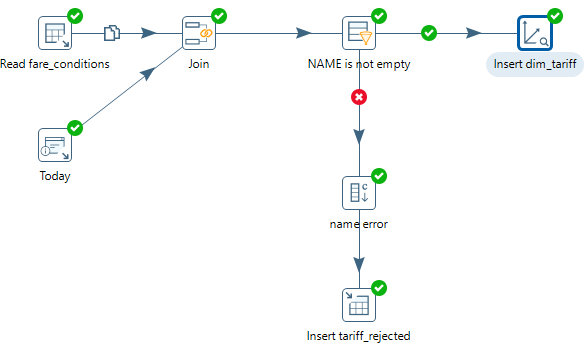
**dim\_tariff**

Источник данных - уникальные значения поля fare\_conditions из таблицы seats, так как в ней содержатся места всех бортов, с указанием класса обслуживания.

Для таблицы dim\_tariff реализованы следующие проверки качества данных:

1. значения в поле airport\_code не могут быть пустыми.

Значения, не прошедшие проверки, записываются в таблицу airports\_rejected.

****

| **Название шага трансформации** | **Тип** | **Описание** |
| --- | --- | --- |
|  | Table input | зачитка данных из таблицы источника bookings.seats |
|  | Get system info | получить текущую дату |
|  | JOIN rows | добавление Today в набор данных |
|  | Filter rows | проверка, что поле name не пустое (NULL) и что его значение не равно пустой строке (“”) |
|  | Add constants | для записей не прошедших проверку добавляем комментарий в поле err\_text о сути ошибки “Name is empty” |
|  | Table output | запись в tariff\_rejected |
|  | Dimension lookup/update | запись данных в хранилище, таблицу dim\_tariff |

**dim\_passengers**

Источник данных - таблица tickets.

Так как все данные о пассажире со временем могут измениться - нет необходимости отслеживать изменения каждой строки. Поэтому измерение dim\_passengers дополняется данными, которых в нем нет по натуральному ключу “passenger\_id - passenger\_name - email - phone”. Для ускорения проверки существования данных в таблице добавлен соответствующий индекс.

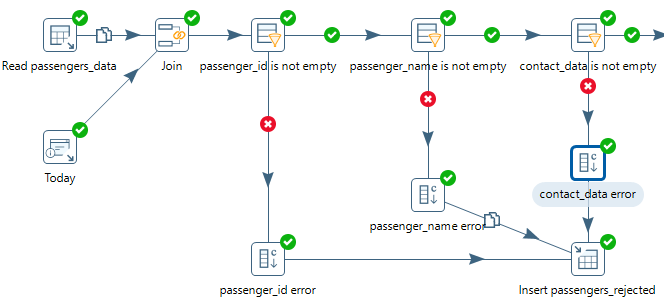
Для таблицы dim\_passengers реализованы следующие проверки качества данных:

1. значения в поле passenger\_id не могут быть пустым;

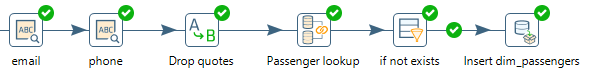
2. значения в поле contact\_data не могут быть пустым;

3. значения в поле passenger\_name не могут быть пустым.

Значения, не прошедшие проверки, записываются в таблицу passengers\_rejected.

****

продолжение



| **Название шага трансформации** | **Тип** | **Описание** |
| --- | --- | --- |
|  | Table input | зачитка данных из таблицы источника bookings.tickets |
|  | Get system info | получить текущую дату |
|  | JOIN rows | добавление Today в набор данных |
|  | Filter rows | проверка, что поле passenger\_id не пустое (NULL) и что его значение не равно пустой строке (“”) |
|  | Add constants | для записей не прошедших проверку добавляем комментарий в поле err\_text о сути ошибки “Passenger id is empty” |
|  | Filter rows | проверка, что поле passenger\_name не пустое (NULL) и что его значение не равно пустой строке (“”) |
|  | Add constants | для записей не прошедших проверку добавляем комментарий в поле err\_text о сути ошибки “Passenger name is empty” |
|  | Filter rows | проверка, что поле contact\_data не пустое (NULL) и что его значение не равно пустой строке (“”) |
|  | Add constants | для записей не прошедших проверку добавляем комментарий в поле err\_text о сути ошибки “Contact data is empty” |
|  | Table output | запись в passengers\_rejected |
|  | Regex evaluation | извлечение адреса электронной почты из поля contact\_data в поле email |
|  | Regex evaluation | извлечение номера телефона из поля contact\_data в поле phone |
|  | Replace in string | удаление из полей email и phone кавычек |
|  | Database join | проверка существования информации о пассажире в измерении dim\_passengers - идентичность по полям passenger\_id, passenger\_name, email, phone  использую database join потому что он позволяет делать более гибкий поиск с помощью секции where sql-запроса, например, обрабатывать null значения |
|  | Filter rows | отсеивание данных, которые уже есть в хранилище |
|  | PostgreSQLbulk loader | запись данных в хранилище, таблицу dim\_passengers |

**fact\_flights**

Источник данных - таблицы tickets, tickets\_flights, flights.

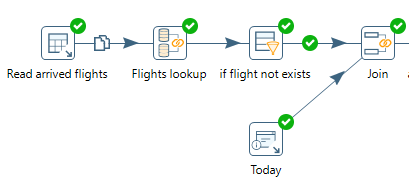
Для таблицы fact\_flights реализованы следующие проверки качества данных:

1. Фактическая дата вылета не может быть больше сегодняшнего дня

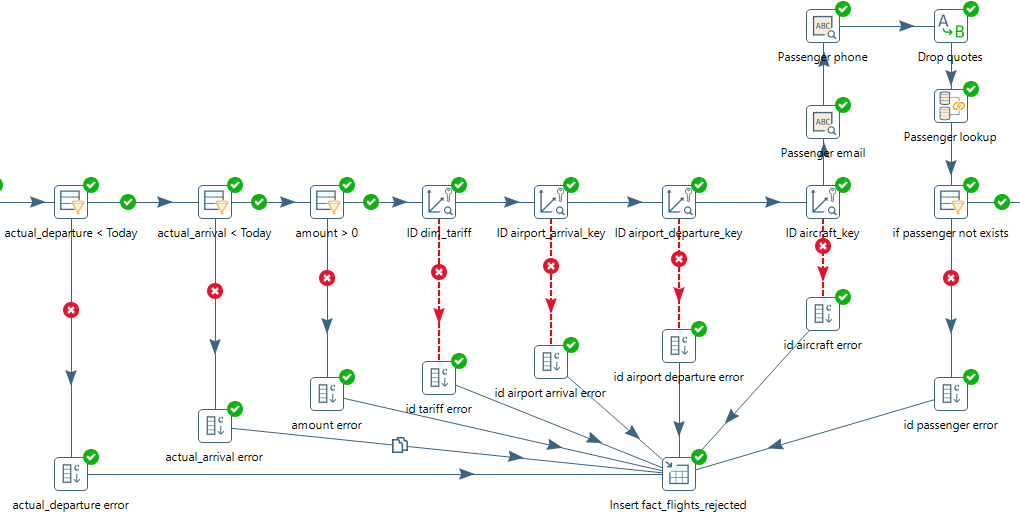
2. Фактическая дата прилета не может быть больше сегодняшнего дня

3. Стоимость должна быть положительной

До пандемии возможна была бы еще проверка, что Аэропорт вылета не равняется Аэропорту прилета.



продолжение



продолжение



| **Название шага трансформации** | **Тип** | **Описание** |
| --- | --- | --- |
|  | Table input | зачитка данных из источника: bookings.flights, bookings.ticket\_flights, bookings.tickets  т.ж. на этом шаге формируем ключи для связи с измерением dim\_calendar, преобразуя фактические даты отправления и прилета в нужный формат YYYYMMDD |
|  | Database join | проверка есть ли уже в хранилище информация по заливаемы данным (по flight\_id) |
|  | Filter rows | если такой flight\_id уже есть, то строку дальше не обрабатываем |
|  | Get system info | получить текущую дату |
|  | JOIN rows | добавление Today в набор данных |
|  | Filter rows | проверка, что фактическая дата вылета меньше текущей даты |
|  | Add constants | для записей не прошедших проверку добавляем комментарий в поле err\_text о сути ошибки “Actual departure later than Today” |
|  | Filter rows | проверка, что фактическая дата прилета меньше текущей даты |
|  | Add constants | для записей не прошедших проверку добавляем комментарий в поле err\_text о сути ошибки “Actual arrival later than Today” |
|  | Filter rows | проверка, что поле amount не пустое (NULL) и что его значение не равно 0.00 |
|  | Add constants | для записей не прошедших проверку добавляем комментарий в поле err\_text о сути ошибки “Amount is empty or less than 0” |
|  | Table output | запись в fact\_flights\_rejected |
|  | Combination lookup / update | поиск ключа в таблице измерения dim\_tariff |
|  | Add constants | если ключ не найден, формирует сообщение об ошибке для записи в fact\_flights\_rejected |
|  | Combination lookup / update | поиск ключа в таблице измерения |
|  | Add constants | если ключ не найден, формирует сообщение об ошибке для записи в fact\_flights\_rejected |
|  | Combination lookup / update | поиск ключа в таблице измерения |
|  | Add constants | если ключ не найден, формирует сообщение об ошибке для записи в fact\_flights\_rejected |
|  | Combination lookup / update | поиск ключа в таблице измерения |
|  | Add constants | если ключ не найден, формирует сообщение об ошибке для записи в fact\_flights\_rejected |
|  | Regex evaluation | извлечение адреса электронной почты из поля contact\_data в поле email |
|  | Regex evaluation | извлечение номера телефона из поля contact\_data в поле phone |
|  | Replace in string | удаление из полей email и phone кавычек |
|  | Database join | проверка существования информации о пассажире в измерении dim\_passengers - идентичность по полям passenger\_id, passenger\_name, email, phone  использую database join потому что он позволяет делать более гибкий поиск с помощью секции where sql-запроса, например, обрабатывать null значения |
|  | Filter rows | если в измерении dim\_passengers данные не найдены, то такие данные не передаются дальше |
|  | Select values | переименование полей с ключами измерений и удаление лишних |
|  | Calculator | расчет задержки времени отправления |
|  | Calculator | расчет задержки времени прибытия |
|  | PostgreSQLbulk loader | запись данных в хранилище, таблицу fact\_flights |