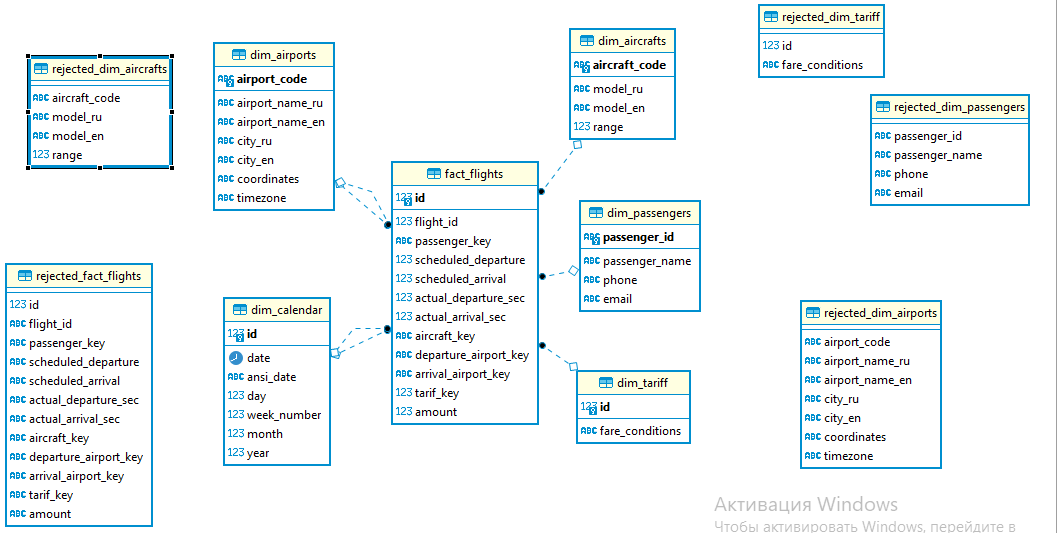
**Описание хранилища данных**

**Диаграмма схемы данных**



**Описание схемы.**

Хранилище данных состоит из таблицы фактов Fact\_Flights, пяти таблиц справочников: Dim\_Airports, Dim\_Aircrafts, Dim\_Passengers, Dim\_Tariff и Dim\_Calendar, а также 5 rejected таблиц для загрузки данных с ошибками в строках.

Основной сущностью хранилища является таблица фактов Fact\_Flights. Данная таблица включает в себя информацию о факте перелета каждого пассажира. В таблицу входят следующие атрибуты: flight\_id, passenger\_key (ссылка на первичный ключ таблицы Dim\_Passengers), scheduled\_departure (ссылка на первичный ключ таблицы Dim\_Calendar), scheduled\_arrival (ссылка на первичный ключ таблицы Dim\_Calendar), actual\_departure\_sec (задержка времени вылета в секундах), actual\_arrival\_sec (задержка времени прилета в секундах), aircraft\_key (ссылка на первичный ключ таблицы Dim\_Aircrafts), departure\_airport\_key (аэропорт вылета, внешний ключ таблицы Dim\_Airports), arrival\_airport\_key (аэропорт прилета, ссылка на первичный ключ таблицы Dim\_Airports), tariff\_key (ссылка на первичный ключ таблицы Dim\_Tariff) и amount (стоимость перелета).

Справочник Dim\_Airports содержит информацию об аэропортах и включает в себя следующие поля: код аэропорта (первичный ключ), наименование аэропорта на английском и русском языках, город, координаты и временная зона.

Справочник Dim\_Aircrafts содержит информацию о самолетах и включает в себя следующие поля: код самолета (первичный ключ), модель самолета на английском и русском языках и максимальная дальность полета каждой модели.

Справочник Dim\_Passengers содержит информацию о фамилии и имени пассажира, номере телефона и электронной почте.

Справочник Dim\_Tariff содержит наименования классов обслуживания.

Справочник Dim\_Calendar содержит сгенерированный ряд дат, состоящий из года, месяца, недели года и дня недели.

**Процедура ETL**

Для выполнения итоговой работы блока DWH был произведен процесс ETL, в котором из облачной базы «bookings» (host: 85.201.153.170; port: 19001; db: demo; user: netology; password: NetoSQL2019) в локальную базу (host: localhost; port: 5432; db: bookings\_dwh; user: postgres; password: \*\*\*\*\*\*\*) были загружены данные.

Для выгрузки данных из облачной базы и загрузки их в необходимом виде и формате был использован инструмент Pentaho Data Integration.

**Справочник дат.**

Для генерации справочника дат были использованы инструменты SQL (скрипт прилагается).

**Справочник тарифов.**

Наполнение справочника тарифов был использован инструмент Pentaho Data Integration. На первом шаге трансформации были выгружены данные из облачной базы (таблица ticket\_flights) с помощью шага трансформации «Table input». Далее с помощью трех шагов трансформации «Filter rows» был проверен столбец fare\_conditions на идентичность строк значениям Economy, Comfort и Business. Строки, в которых могут содержаться ошибочные данные на шаге «Filter rows» загружаются в таблицу rejected\_dim\_tariff.

**Справочник аэропортов**.

Для наполнения справочника аэропортов был использован инструмент Pentaho Data Integration. На первом шаге трансформации были выгружены данные из облачной базы (таблица airports\_data) с помощью шага трансформации «Table input». Далее с помощью трех шагов трансформации «Filter rows» были проверены столбцы airport\_code, name\_ru и city\_ru на отсутствие null значений в строках. Строки, в которых могут содержаться ошибочные данные на шаге «Filter rows» загружаются в таблицу rejected\_dim\_airports.

**Справочник самолетов**.

Для наполнения справочника самолетов на первом шаге трансформации были выгружены данные из облачной базы (таблица aircrafts\_data) с помощью шага трансформации «Table input». С помощью шага трансформации «Filter rows» были проверены строки в столбце «aircraft\_code» на отсутствие в строках значений null. На следующем шаге трансформации «Filter rows» была произведена проверка столбца «range» на сравнение показателей дальности полета и 0 (range > 0). Строки, в которых могут содержаться ошибочные данные на шаге «Filter rows» загружаются в таблицу rejected\_dim\_aircrafts.

**Справочник пассажиров**.

Для наполнения справочника пассажиров на первом шаге трансформации были выгружены данные из облачной базы (таблицы tickets) с помощью шага трансформации «Table input». С помощью шага трансформации «Filter rows» были проверены строки в столбце «passenger\_name» на отсутствие в строках значений null. На следующем шаге трансформации «Filter rows» была произведена проверка столбца «passenger\_id» на отсутствие в строках значений null. На следующем шаге трансформации «Filter rows» была произведена проверка двух столбцов «email» и «phone» на отсутствие значений null в одном или обеих столбцах одной строки. Строки, в которых могут содержаться ошибочные данные на шаге «Filter rows» загружаются в таблицу rejected\_dim\_passenger.

**Таблица фактов.**

Для наполнения таблицы фактов на первом шаге трансформации были выгружены данные из облачной базы (таблица tickets, ticket\_flights, flights, aircrafts\_data и airports\_data) с помощью шага трансформации «Table input». С помощью шага трансформации «Filter rows» были отфильтрованы строки со значением «Arrived», остальные строки были направлены в шаг трансформации «Dummy (do nothing)».

На следующем шаге трансформации был использован калькулятор для создания новых столбцов actual\_departure\_sec (задержка времени вылета в секундах), actual\_arrival\_sec (задержка времени прилета в секундах). На этом шаге был произведен подсчет на основе данных из столбцов actual\_departure, scheduled\_departure, actual\_arrival и scheduled\_arrival. От фактического времени вылета вычитается планируемое время вылета и результат отображаем в секундах. От фактического времени прилета вычитается планируемое время прилета и результат отображаем в секундах.

На следующем шаге трансформации «Replace in string» мы меняем значения в столбце «fare\_conditions» с названия класса обслуживания (Economy, Comfort и Business) на нумерацию, соответствующую уже имеющимся в таблице «Dim\_Tariff».

На следующем шаге трансформации «Select values» мы во вкладке Remove удаляем ненужные столбцы «Status», «actual\_arrival» и «actual\_departure», которые использовались для фильтрации и подсчета необходимых нам значений в калькуляторе. Также на этом этапе было произведено изменение метаданных (вкладка Meta-data) в столбцах scheduled\_arrival (на тип данных string в формате «YYYYMMDD»), scheduled\_departure (на тип данных string в формате «YYYYMMDD») и fare\_conditions (на тип данных integer).

На следующем шаге трансформации «Filter rows» была произведена проверка столбца «passenger\_id» на отсутствие в строках значений null. Строки, в которых могут содержаться ошибочные данные на шаге «Filter rows» загружаются в таблицу rejected\_fact\_flight.

На следующем шаге трансформации «Select values 2» мы меняем метаданные значений столбцов scheduled\_arrival, scheduled\_departure и amount на тип integer.

**Файлы, прилагаемые к данной работе**

aircraft dwh - трансформация для наполнения таблицы Dim\_Aircraft

dwh airports - трансформация для наполнения таблицы Dim\_Airports

dwh flights - трансформация для наполнения таблицы Fact\_Flights

dwh passenger - трансформация для наполнения таблицы Dim\_Passenger

aircraft transformation - скриншот трансформации для наполнения таблицы Dim\_Aircraft

airports transformation - скриншот трансформации для наполнения таблицы Dim\_Airports

flight fact transformation - скриншот трансформации для наполнения таблицы Fact\_Flights

passenger transformation - скриншот трансформации для наполнения таблицы Dim\_Passenger

dwh itog - скрипт создания таблиц хранилища данных