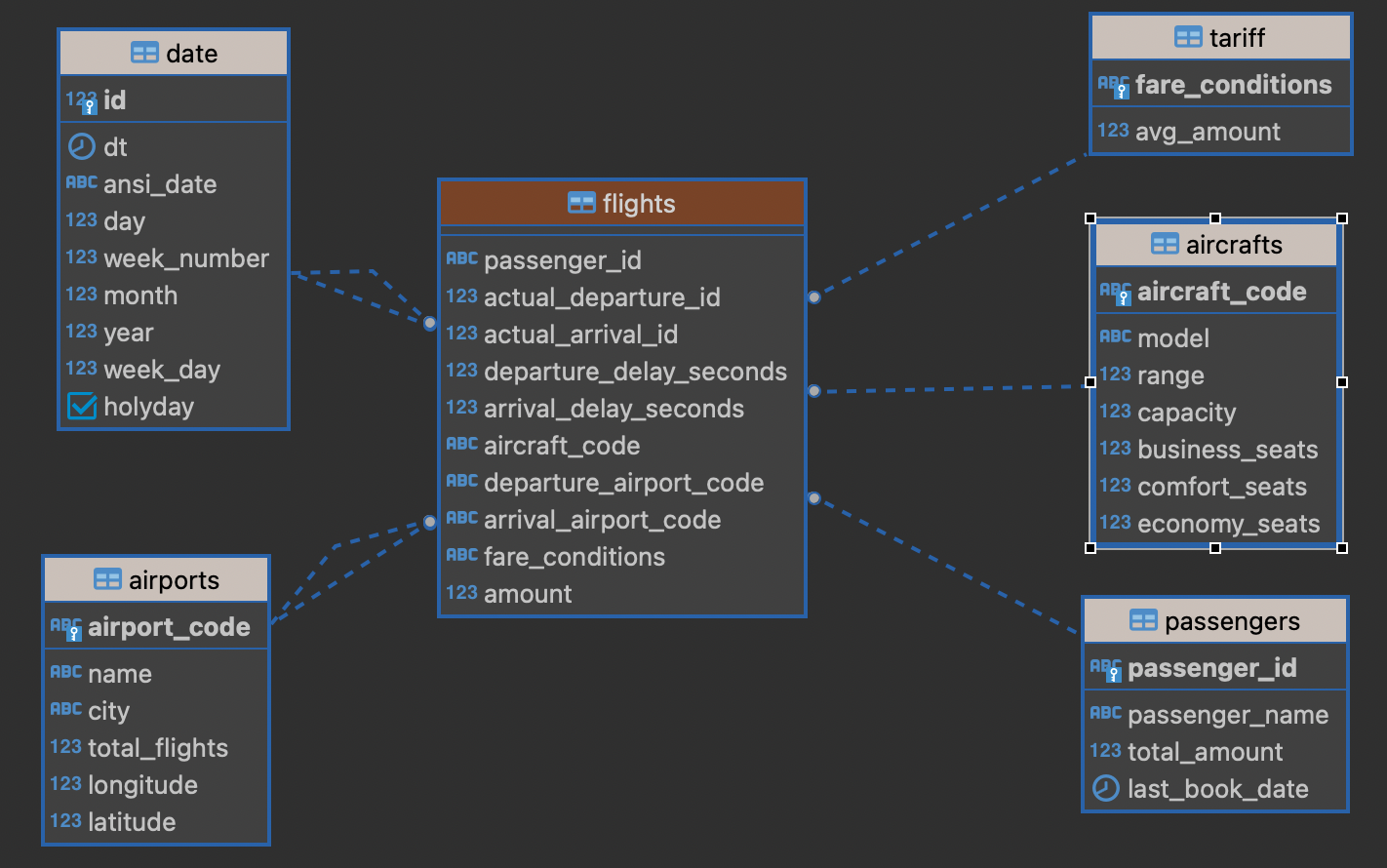
**Проектная работа по модулю**

**“DWH”**



В ходе работы спроектирована и реализована витрина данных по авиаперевозкам на основе данных базы данных bookings построенная по схеме звезда. Витрина состоит из 5 таблиц-измерений и таблицы фактов. Описание таблиц представлено в таблице А.1 приложения А. Описание процессов ETL проверки качества данных и их для загрузки представлено приложение Б. Исходные скрипты создания таблиц витрины и скрипты ETL процессов разработанных в Pentaho DI доступны по ссылке:

Github:

**Приложение А**

Таблица А.1 - Описание таблиц представлено

|  |  |  |  |
| --- | --- | --- | --- |
| **Наименование таблицы** | **Столбцы таблицы** | **Описание столбцов** | **Примечание** |
| Таблицы измерений витрины | | | |
| dim.date | id  dt date  ansi\_date  day int  week\_number  month int  year int  week\_day  holyday | дата в типе int  дата  дата в типе char  день месяца  Номер недели в году  Номер месяца в году  Год  День недели  Выходной день или нет | Первичный ключ |
| dim.passengers | passenger\_id  passenger\_name  total\_amount  last\_book | id пассажира  имя пассажира  общая сумма бронирования  последняя дата бронирования | Первичный ключ |
| dim.aircrafts | aircraft\_code  model  range  capacity  business\_seats  comfort\_seats  economy\_seats | уникальный код  модель самолета  дальность полета  вместимость  количество мест бизнес-класса  количество мест комфорт класса  количество мест эконом класса | Первичный ключ |
| dim.airports | airport\_code  name  city  total\_flights longitude  latitude | уникальный код  имя  город  всего полетов  долгота  широта | Первичный ключ |
| dim.tariff | fare\_conditions  avg\_amount | тариф  средняя цена | Первичный ключ |
| Таблица фактов | | | |
| fact.flights | passenger\_id  actual\_departure\_id  actual\_arrival\_id  departure\_delay\_seconds  arrival\_delay\_seconds  aircraft\_code  departure\_airport\_code  arrival\_airport\_code  fare\_conditions  amount | id пассажира  дата и время вылета (факт)  дата и время прилета (факт)  задержка вылета в секундах  задержка прилета в секундах  код самолета  код аэропорта вылета  код аэропорта прилета  класс обслуживания  стоимость | вешний ключ к dim.passengers  вешний ключ к dim.date  вешний ключ к dim.date  вешний ключ к dim.aircrafts  вешний ключ к dim.airports  вешний ключ к dim.airports  вешний ключ к dim.tariff |

**Приложение Б**

Таблица Б.1 - Описание процессов ETL

|  |  |  |
| --- | --- | --- |
| **Имя ETL процесса** | **Описание ETL процесса** | **Реализованные проверки** |
| dim\_calendar | Генерирует диапазон дат между минимальной датой в базе данных (БД) и текущей датой. Проверяет при помощи внешнего API сервиса являют ли даты выходными днями и преобразовывает даты в колонки необходимые для заполнения таблицы dim.date. После данные загружаются в таблицу dim.date и в csv файл. | - |
| dim\_passengers | Собирает данные о пассажирах и их бронировании из таблиц bookings.tickets и bookings.booking. Находит дату последнего бронирования и проверяет качество данных. После данные загружаются в таблицу dim.passengers и в csv файл. Непрошедшие проверку данные загружаются 'rejected’ csv файл. | проверка на дубликаты  проверка на не нулевую сумму брони  проверка на отрицательную сумму брони |
| dim\_aircrafts | Собирает данные о самолетах и местах в самолете из таблиц bookings.aircrafts и bookings.seats. Подсчитывает количеством мест каждого класса в самолетах. Вычисляет общую вместимость самолета и проверяет качество данных. После данные загружаются в таблицу dim.aircrafts и в csv файл. Непрошедшие проверку данные загружаются 'rejected’ csv файл. | проверка на наличие модели самолета  проверка на ненулевую вместимость  проверка на расстояние полета больше 0 |
| dim\_airports | Собирает данные об аэропортах, общем количестве полетов из данного аэропортов из таблицы bookings.flights. Проверяет качество данных и после данные загружаются в таблицу dim.airports и в csv файл. Непрошедшие проверку данные загружаются 'rejected’ csv файл. | проверка на наличие имени аэропорта  проверка на наличие города  проверка на количества полетов больше 0 |
| dim\_tariff | Собирает данные о тарифах, их средней стоимости из bookings.ticket\_flights. Проверяет качество данных и после данные загружаются в таблицу dim.tariff и в csv файл. Непрошедшие проверку данные загружаются 'rejected’ csv файл. | проверка на наличие имени тарифа  проверка на средняя стоимость не null  проверка на средняя стоимость больше 0 |
| fact.flights | Cобирает необходимы данные для таблицы fact\_flights. Проверяет качество данных и после данные загружаются в таблицу fact\_flights. и в csv файл. Непрошедшие проверку данные загружаются 'rejected’ csv файл. | проверка на код пасажира не null  проверка на стоимость не null  проверка на стоимость больше 0 |