**Дипломный проект по профессии Инженер данных**

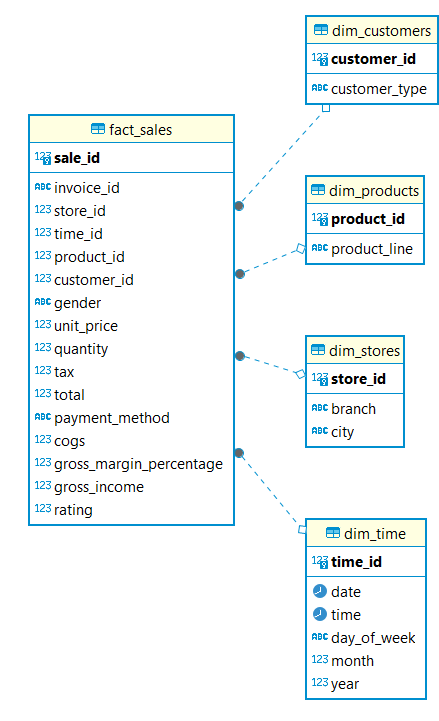
# 1. Документированная схема хранилища данных

**Описание:**  
Хранилище данных (DWH) структурировано по **звездной схеме** (Star Schema), включающей **факт** и **измерения**.

**Таблицы хранилища:**

1. **Факт продаж (fact\_sales)**
   * sale\_id (PK) – идентификатор продажи
   * invoice\_id (FK) – Номер заказа
   * store\_id (FK) – идентификатор магазина
   * time\_id – идентификатор даты
   * customer\_id (FK) – клиент
   * product\_id (FK) – товар
   * gender – Пол покупателя
   * unit\_price – цена за шт
   * quantity - количество
   * tax -налог
   * total - сумма
   * payment\_method – метод оплаты
   * cogs - себестоимость
   * gross\_margin\_percentage – валовый процент маржинальности
   * gross\_income – валовая прибыль
   * rating - рейтинг
2. **Измерение клиентов (dim\_customers)**
   * customer\_id (PK) – ID клиента
   * customer\_type – Тип клиента
3. **Измерение товаров (dim\_products)**
   * product\_id (PK) – ID товара
   * product\_line – Категория товара
4. **Измерение магазинов (dim\_stores)**
   * store\_id (PK) – ID магазина
   * branch – Отрасль
   * city – Город
5. **Измерение дат (dim\_dates)**
   * time\_id (PK) – ID даты
   * date – Дата
   * year – Год
   * month – Месяц
   * day\_of\_week– День недели
   * time – Время

ERD-диаграмма:



# 2. Документированная схема ETL-процессов

Описание:  
ETL состоит из **трех этапов**:

* **Extract (извлечение)** – загрузка данных из исходных систем
* **Transform (преобразование)** – очистка, обогащение, агрегация
* **Load (загрузка)** – загрузка в DWH

## **Шаги ETL:**

**Извлечение (Extract)**

* Источники: файлы CSV, PostgreSQL
* Данные загружаются во временные таблицы

**Преобразование (Transform)**

* Очистка данных (удаление дубликатов, NULL-значений)
* Приведение типов данных (например, date → timestamp)
* Создание date\_id (даты в числовом формате YYYYMMDD)
* Агрегация продаж по product\_id

**Загрузка (Load)**

* Загрузка данных в NDS и DDS слой
* Измерения обновляются инкрементально

## Схема ETL-процесса:

