

Dokumentacja „Geeks & Dragons”

Urszula Spik, Natalia Jelito, Maciej Karczewski, Magdalena Szymkowiak

2023-06-30

1 Spis użytych technologii

Do stworzenia projektu zostały użyte następujące technologie:

1. Tworzenie bazy danych:
 - Python z bibliotekami:
 - pandas,
 - numpy,
 - random,
 - datetime,
 - urllib.parse,
 - sqlalchemy,
 - itertools,
 - faker,
 - MariaDB,
2. Analiza danych:
 - R z bibliotekami:
 - RMariaDB,
 - ggplot2,
 - dplyr,
 - dbplyr,
 - tidyr,
 - kableExtra,
 - emojiFont,
 - stringr,
 - colorspace,
 - wesanderson.
3. Raport i dokumentacja:
 - Rmarkdown
4. Tworzenie schematu baz danych:
 - Miro

2 Lista plików i opis ich zawartości

W folderze **bazy_danych** znajdują się pliki tworzące cały projekt. Struktura zawartości folderu wygląda następująco

```
bazy_danych
├── data
│   ├── address.csv
│   ├── english_first_name.csv
│   ├── english_last_name.csv
│   └── games.csv
```

```

├── polish_female_last_name.csv
├── polish_female_name.csv
├── polish_male_last_name.csv
├── polish_female_name.csv
├── dokumentacja.pdf
├── Raport
│   ├── raport.Rmd
│   └── raport.pdf
├── scr
│   ├── __init__.py
│   ├── create.sql
│   └── generate.py
└── main.py

```

W folderze *data* znajdują się pliki *.csv*, których użyliśmy do generowania adresów, imion, nazwisk zarówno klientów sklepu jak i też samych pracowników. Użyliśmy również ich do generowania gier, które można wypożyczyć albo kupić. Źródła danych:

- *address.csv* - <https://bip.um.wroc.pl/artykul/389/50007/aktualne-wykazy-nazw-ulic-i-adresow>
- *english_first_name.csv* - <https://catalog.data.gov/dataset/popular-baby-names/resource/02e8f55e-2157-4cb2-961a-2aabb75cbc8b>
- *english_last_name.csv* - https://www.census.gov/topics/population/genealogy/data/2010_surnames.html
- *games.csv* - <https://www.kaggle.com/datasets/andrewmvd/board-games>
- *polish_female_last_name.csv* oraz *polish_male_last_name.csv* - <https://dane.gov.pl/pl/dataset/1681,nazwiska-osob-zyjacych-wystepujace-w-rejestrze-pesel>
- *polish_female_name.csv* oraz *polish_male_name.csv* - <https://dane.gov.pl/pl/dataset/219,imi-ona-nadawane-dzieciom-w-polsce>

Plik *dokumentacja.pdf* jest plikiem zawierającym dokumentację.

W folderze *Raport* znajdują się plik do generowania raportu oraz plik z gotowym już raportem.

Kolejnym folderem jest *scr*. W środku znajdują się pliki, które służą do generowania i tworzenia bazy danych. Plik *__init__.py*, potrzebny jest do możliwości importowania pythonowych funkcji z innych plików znajdujących się w tym samym folderze. Plik *create.sql* tworzy nam tabele do bazy danych. Plik *generate.py* zawiera funkcje, przy pomocy których generujemy dane do wypełnienia stworzonej wcześniej bazy danych.

Plik *main.py* służy do wygenerowania tabel z danymi oraz wgrania ich na serwer.

3 Kolejność i sposób uruchamiania plików, aby uzyskać gotowy projekt

1. Posiadaj Pythona 3.7.9
2. Zainstaluj potrzebne pakiety komendą: ``pip install -r requirements.txt``
3. Uruchom program *main.py*: ``python main.py``
4. Uruchom i skompiluj za pomocą Knit plik ``report.rmd``

4 Schemat bazy danych i zależności funkcyjne w tabelach

Klucze główne zaznaczamy przez pogrubienie.

4.1 Tabela Customers

R(**customer_id**, first_name, last_name, birth_date, email, phone)

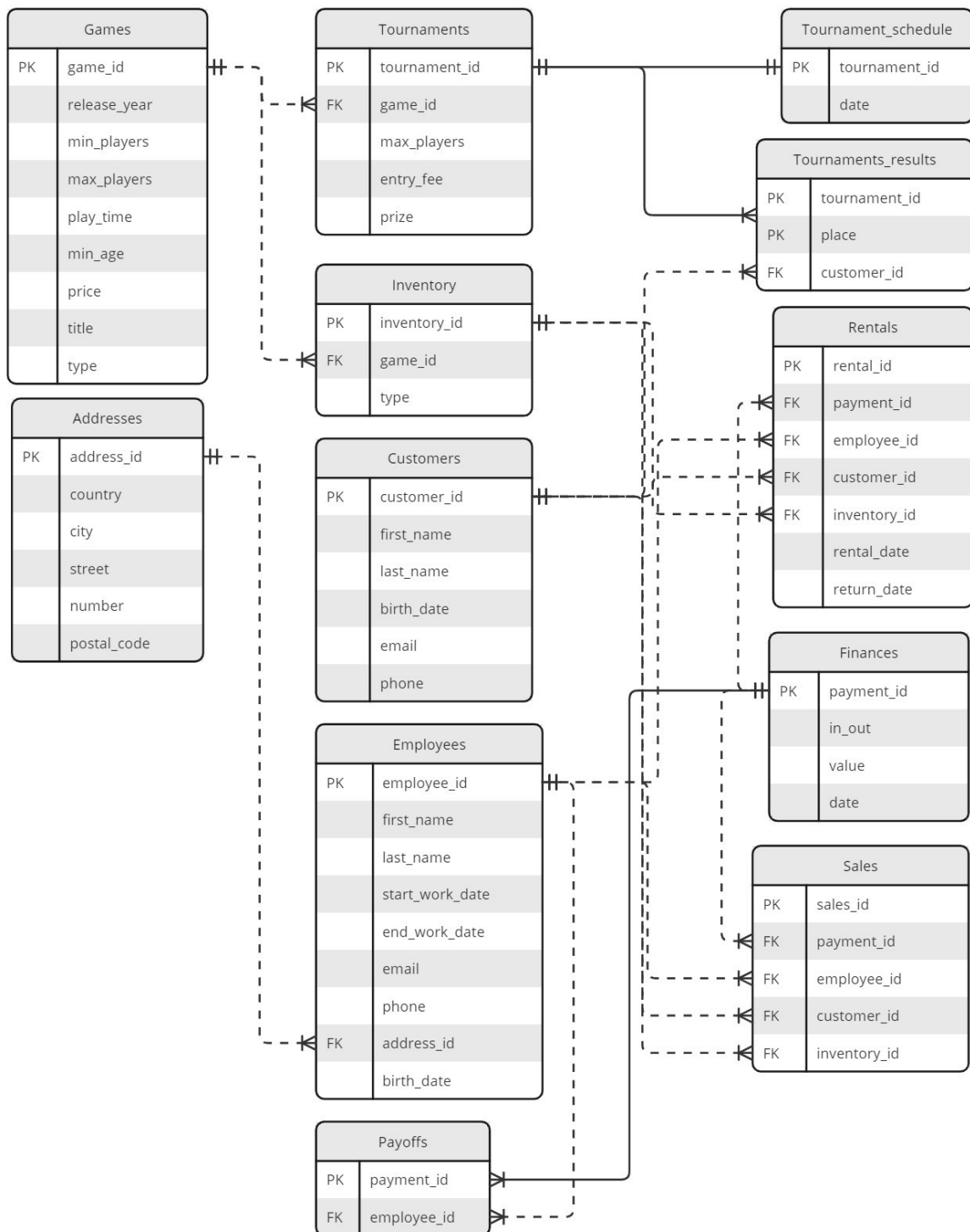


Figure 1: Schemat bazy danych

- customer_id [SMALLINT UNSIGNED] : numer identyfikacyjny klienta
- first_name [VARCHAR(45)]: imię klienta
- last_name [VARCHAR(50)]: nazwisko klienta
- birth_date [DATE]: data urodzenia klienta
- email [VARCHAR(50)]: email klienta
- phone [VARCHAR(12)]: numer telefonu klienta

$$\Sigma = \{ \text{customer_id} \rightarrow \text{first_name}, \text{customer_id} \rightarrow \text{last_name}, \text{customer_id} \rightarrow \text{birth_date}, \\ \text{customer_id} \rightarrow \text{email}, \text{email} \rightarrow \text{customer_id}, \text{customer_id} \rightarrow \text{phone}, \text{phone} \rightarrow \text{customer_id} \}$$

Informacje o dokładnym kliencie możemy pozyskać przez customer_id. Dokładną identyfikację klienta zapewnia także jego email lub jego numer telefonu, ponieważ zakładamy, że każdy ma jeden unikalny numer telefonu oraz adres email.

4.2 Tabela Employees

R(**employee_id**, first_name, last_name, start_work_date, end_work_date, email, phone, address_id)

- employee_id [SMALLINT UNSIGNED]: numer identyfikacyjny pracownika
- first_name [VARCHAR(45)]: imię pracownika
- last_name [VARCHAR(50)]: nazwisko pracownika
- start_work_date [DATE]: data zatrudnienia (rozpoczęcie pracy)
- end_work_date [DATE]: data zakończenia pracy
- email [VARCHAR(50)]: adres email pracownika
- phone [VARCHAR(12)]: numer telefonu pracownika
- address_id [SMALLINT UNSIGNED] (klucz obcy odwołujący się do tabeli Addresses : numer identyfikacyjny adresu).

$$\Sigma = \{ \text{employee_id} \rightarrow \text{first_name}, \text{employee_id} \rightarrow \text{last_name}, \text{employee_id} \rightarrow \text{start_work_date}, \\ \text{employee_id} \rightarrow \text{start_work_date}, \text{employee_id} \rightarrow \text{email}, \text{email} \rightarrow \text{employee_id}, \text{employee_id} \rightarrow \text{phone}, \text{phone} \rightarrow \text{employee_id} \}$$

Pracownik jest jednoznacznie zdefiniowany przez employee_id. Tak jak w przypadku klientów, pracownika też możemy rozpoznać po jego numerze telefonu lub email.

4.3 Tabela Addresses

R(**address_id**, country, city, street, number, postal_code)

- address_id [SMALLINT UNSIGNED]: numer identyfikacyjny adresu.
- country [VARCHAR(50)]: państwo
- city [VARCHAR(50)]: miasto
- street [VARCHAR(50)]: ulica
- number [VARCHAR(10)]: numer budynku
- postal_code [VARCHAR(10)]: kod pocztowy.

$$\Sigma = \{ \text{address_id} \rightarrow \text{country}, \text{address_id} \rightarrow \text{city}, \text{address_id} \rightarrow \text{street}, \text{address_id} \rightarrow \text{number}, \\ \text{address_id} \rightarrow \text{postal_code}, \text{country} + \text{city} + \text{street} + \text{number} + \text{postal_code} \rightarrow \text{address_id} \}$$

Adres jest dokładnie, jednoznacznie wyznaczony dla danego address_id. Pozostałe pola tylko razem wskażą nam dokładny adres.

4.4 Tabela Games:

R(game_id, title, release_year, min_players, max_players, play_time, min_age, type, price)

- game_id [SMALLINT UNSIGNED]: numer identyfikacyjny gry
- title [VARCHAR(100)]: oryginalny tytuł gry
- release_year [DATE]: rok wydania gry
- min_players [TINYINT UNSIGNED]: minimalna liczba graczy
- max_players [TINYINT UNSIGNED]: maksymalna liczba graczy
- play_time [SMALLINT UNSIGNED]: szacowany czas gry w minutach
- min_age [TINYINT UNSIGNED]: wiek minimalny do zagrania w grę
- type [VARCHAR(18)]: typ gry tzn czy jest to gr czy rozszerzenie do gry
- price [DECIMAL]: cena za kupno gry

$$\Sigma = \{\text{game_id} \rightarrow \text{title}, \text{game_id} \rightarrow \text{release_year}, \text{game_id} \rightarrow \text{min_players}, \text{game_id} \rightarrow \text{max_players}, \\ \text{game_id} \rightarrow \text{play_time}, \text{game_id} \rightarrow \text{min_age}, \text{game_id} \rightarrow \text{type}, \text{game_id} \rightarrow \text{price}\}$$

Gra jest wyznaczona jednoznacznie przez game_id. Połączenie wszystkich pozostałych informacji nie musi nam jednoznacznie określić gry (Tytuły nie muszą być unikatowe)

4.5 Tabela Inventory

R(inventory_id, game_id, type)

- inventory_id [SMALLINT UNSIGNED]: numer identyfikacyjny inwentarza
- game_id [SMALLINT UNSIGNED] (klucz obcy odwołujący się do tabeli Games): numer identyfikacyjny grę
- type [CHAR(1)]: przeznaczenie gry na magazynie (czy na sprzedaż, czy na wypożyczenie, czy na turniej)

$$\Sigma = \{\text{inventory_id} \rightarrow \text{game_id}, \text{inventory_id} \rightarrow \text{type}\}$$

Gra w magazynie jest wyznaczona jednoznacznie przez inventory_id. Reszta informacji nie pozwala nam tego określić ponieważ możemy mieć kilka takich samych tytułów o takim samym przeznaczeniu.

4.6 Tabela Tournaments

R(tournament_id, game_id, max_players, entry_fee, prize)

- tournament_id [SMALLINT UNSIGNED]: numer identyfikacyjny turnieju
- game_id [SMALLINT UNSIGNED] (klucz obcy odwołujący się do tabeli Games): numer identyfikacyjny grę w jaką się gra na turnieju
- max_players [TINYINT UNSIGNED]: maksymalna liczba graczy na turnieju
- entry_fee [DECIMAL]: wpisowe
- prize [DECIMAL]: nagroda za turniej zamiast

$$\Sigma = \{\text{tournament_id} \rightarrow \text{game_id}, \text{tournament_id} \rightarrow \text{max_players}, \\ \text{tournament_id} \rightarrow \text{entry_fee}, \text{tournament_id} \rightarrow \text{prize}\}$$

Turniej jest jednoznacznie wyznaczony jedynie przez tournament_id. Zakładamy, że nagroda nie jest bezpośrednio zależna od wpisowego i liczby graczy.

4.7 Tabela Tournament_schedule

R(tournament_id, date)

- tournament_id [SMALLINT UNSIGNED] (klucz obcy odwołujący się do tabeli Tournaments): numer identyfikacyjny turnieju
- date [DATE]: data turnieju

$$\Sigma = \{\text{tournament_id} \rightarrow \text{date}\}$$

Tabela łącząca turniej z jego datą. W jeden dzień może odbywać się więcej niż 1 turniej

4.8 Tabela Tournaments_results

R(tournament_id, place, customer_id)

- tournament_id [SMALLINT UNSIGNED] (klucz obcy odwołujący się do tabeli Tournaments): numer identyfikacyjny turnieju
- place [SMALLINT UNSIGNED]: miejsce zajęte w turnieju przez gracza (klienta)
- customer_id [SMALLINT UNSIGNED] (klucz obcy odwołujący się do tabeli Customers): numer identyfikacyjny klienta

$$\Sigma = \{\text{tournament_id} + \text{place} \rightarrow \text{client_id}\}$$

Identyfikator turnieju w połączeniu z miejscem wyznacza nam jednoznacznie osobę, która te miejsce zajęła w turnieju. Zakładamy, że w turniejach nie ma miejsc ex aequo.

4.9 Tabela Finances

R(payment_id, in_out, value, date)

- payment_id [MEDIUMINT UNSIGNED]: numer identyfikacyjny płatności
- in_out [TINYINT]: czy jest to płatność czy przychód
- value [DECIMAL]: wartość płatności
- date [DATE]: data płatności

$$\Sigma = \{\text{payment_id} \rightarrow \text{in_out}, \text{payment_id} \rightarrow \text{value}, \text{payment_id} \rightarrow \text{date}\}$$

Tylko payment id wyznacza nam jednoznacznie przepływ pieniężny.

4.10 Tabela Payoffs: R(payment_id, employee_id)

- payment_id [MEDIUMINT UNSIGNED] (klucz obcy odwołujący się do tabeli Finances): numer identyfikacyjny płatności
- employee_id [SMALLINT UNSIGNED] (klucz obcy odwołujący się do tabeli Employees): numer identyfikacyjny pracownika

$$\Sigma = \{\text{payment_id} \rightarrow \text{employee_id}\}$$

Każda wypłata jest jednoznacznie zdefiniowana przez payment_id i ma przypisanego pracownika. Relacją w drugą stronę nie zachodzi ponieważ pracownik może mieć kilka wypłat.

4.11 Tabela Rentals

R(**rental_id**, payment_id, employee_id, customer_id, inventory_id, rental_date, return_date)

- rental_id [SMALLINT UNSIGNED]: numer identyfikacyjny wypożyczenia
- payment_id [MEDIUMINT UNSIGNED] (klucz obcy odwołujący się do tabeli Finances) : numer identyfikacyjny płatności
- employee_id [SMALLINT UNSIGNED] (klucz obcy odwołujący się do tabeli Employees): numer identyfikacyjny pracownika
- customer_id [SMALLINT UNSIGNED] (klucz obcy odwołujący się do tabeli Customers): numer identyfikacyjny klienta
- inventory_id [SMALLINT UNSIGNED] (klucz obcy odwołujący się do tabeli Inventory): numer identyfikacyjny inwentarza
- rental_date [DATE]: data wypożyczenia gry
- return_date [DATE]: data oddania gry

$$\Sigma = \{\text{rental_id} \rightarrow \text{payment_id}, \text{rental_id} \rightarrow \text{employee_id}, \text{rental_id} \rightarrow \text{customer_id}, \text{rental_id} \rightarrow \text{inventory_id}, \\ \text{rental_id} \rightarrow \text{rental_date}, \text{rental_id} \rightarrow \text{return_date}, \text{payment_id} \rightarrow \text{rental_id}\}$$

Wypożyczenie jest jednoznacznie wyznaczone przez rental_id, tak jak payment id. Pozostałe pola nie muszą jednoznacznie wyznaczyć wypożyczenia.

4.12 Tabela Sales

R(**sales_id**, payment_id, employee_id, customer_id, inventory_id)

- sales_id [SMALLINT UNSIGNED]: numer identyfikacyjny sprzedaży gry
- payment_id [MEDIUMINT UNSIGNED] (klucz obcy odwołujący się do tabeli Finances) : numer identyfikacyjny płatności
- employee_id [SMALLINT UNSIGNED] (klucz obcy odwołujący się do tabeli Employees): numer identyfikacyjny pracownika
- customer_id [SMALLINT UNSIGNED] (klucz obcy odwołujący się do tabeli Customers): numer identyfikacyjny klienta
- inventory_id [SMALLINT UNSIGNED] (klucz obcy odwołujący się do tabeli Inventory): numer identyfikacyjny inwentarza

$$\Sigma = \{\text{sales_id} \rightarrow \text{payment_id}, \text{sales_id} \rightarrow \text{employee_id}, \text{sales_id} \rightarrow \text{customer_id}, \\ \text{sales_id} \rightarrow \text{game_id}, \text{payment_id} \rightarrow \text{sales_id}\}$$

Każda sprzedaż ma unikatowy sales_id. Transakcja sprzedaży może być też jednoznacznie wyznaczona przez payment_id.

5 Postać normalna bazy danych

Jak możemy zauważyć w naszej bazie danych w każdej tabeli wszystkie nietrywialne relacje zaczynają się od nadklucza więc spełnia postać BCNF z czego wynika także że spełnia postać normalną EKNF.

6 Co było najtrudniejsze w projekcie?

Największym wyzwaniem w projekcie było poprawne wygenerowanie danych i wrzucenie ich na serwer. Wyzwaniem było także zrozumienie i zaprojektowanie bazy danych w ten sposób by spełniała postać normalną EKNF.