Report Project Databases: Game Industry

Kélian PONS, Clément MARTIN
1st year Master students, Master Data Science
Université de Lille

December 7, 2025

Contents

Introduction

The aim of the project for the Databases course is to create a database from a chosen data set. For this project we choose to create a database for the video game industry, with information on the games, the developers, the country and town of their headquarters.

1 Data of the project

We started by downloading data from different sources:

- Video game studios: From Kaggle, download here
- Games: From Huggingface, download here
- World cities: From Kaggle, download here
- World countries: From Github, download here

1.1 Modification on Video game studios dataset

The video game studios dataset is a dataset from Kaggle, which contains information on the video game studios. The csv file contains the following atributes:

- **Developer**: name of the video game studio
- **City**: city where the video game studio is located
- **Administrative division**: administrative division of the city where the video game studio is located
- **Country**: country where the video game studio is located
- **Est.**: year when the video game studio was founded
- **Notable games, series or franchises**: list of notable games, series or franchises of the video game studio
- **Notes**: notes about the video game studio

Frist manual cleaning (87 lines) was done on video-games-developers.csv to match the other datasets. Especially the **City** and **Country** columns, which contain non standard names. You can find all the manual modification in the modif_video_game_studio.json file. We also add a new column called **DeveloperID** to be the primary key of the table.

Example of missmatched data between game studios and cities datasets

```
Ultimate Play the Game,"Ashby-de-la-Zouch",England,United Kingdom,1982,JetpacLunar JetmanAtic AtacSabreman
```

Video game studios dataset

```
"Ashby de la Zouch","Ashby de la Zouch","52.7460","-1.4760","United Kingdom","GB","GBR","Leicestershire","",
```

World cities dataset

Then a DeveloperID base on the Developer column was created to be the primary key of the table. It will also be the foreign key of the games table. The City column was modified to CityID to match the cities table and be the foreign key. The Country column was also modified to CountryID and it will be the foreign key of the countries table.

1.2 Modification on the Games dataset

The games dataset is a dataset from Huggingface, which contains information on the games. The csv file contains 40 atributes : AppID, Name, Release date, Estimated owners, Peak CCU, Required age, Price, Discount, DLCcount, About the game, Supported languages, Full audio languages, Reviews, Header image, Website, Support url, Support email, Windows, Mac, Linux, Metacritic score, Metacritic url, User score, Positive, Negative, Score rank, Achievements, Recommendations, Notes, Average playtime forever, Average playtime two weeks, Median playtime forever, Median playtime two weeks, Developers, Publishers, Categories, Genres, Tags, Screenshots, Movies,
Only the following atributes were kept :

- **AppID**: The ID of the game.
- **Name**: The name of the game.
- **Release date**: The release date of the game.
- **Estimated owners**: The estimated range of owners of the game.
- **Required age**: The required age to play the game.
- **Price**: The price of the game.
- **DLCcount**: The number of DLCs in the game.
- **Supported languages**: The languages supported by the game.
- **Windows**: True if the game is supported on Windows, False otherwise.
- **Mac**: True if the game is supported on Mac, False otherwise.
- **Linux**: True if the game is supported on Linux, False otherwise.
- **Metacritic score**: The metacritic score of the game.
- **User score**: The user score of the game.
- **Positive**: The number of positive reviews of the game.
- **Negative**: The number of negative reviews of the game.
- **Achievements**: Number of achievements in the game.
- **Average playtime forever**: The average playtime of the game.
- **Developers**: The developers of the game.
- **Categories**: The categories of the game.
- **Genres**: The genres of the game.

We dropped the rows that have missing values for the Developers. The Developers column was modified to DeveloperID to match the DeveloperID column in the video game studios dataset.

Since game can have multiple genres, the genres were split into additional csv files. One file contains the genresID and name of the genres, and the other file contains the link between the AppID and the genresID, since the game can have multiple genres one AppID can appear multiple in rows. The same thing was done for the categories and languages supported.

1.3 Modification on the World cities dataset

The world cities dataset is a dataset from Kaggle, which contains information on the world cities. The csv file contains the following atributes:

- **id**: id of the city.
- **City**: name of the city.
- **City ASCII**: name of the city in ASCII.
- **Latitude**: latitude of the city.
- **Longitude**: longitude of the city.
- **Country**: country of the city.
- **ISO2**: ISO2 code of the country.
- **ISO3**: ISO3 code of the country.
- **Admin name**: administrative name of the city.
- **Capital**: Information on the city as a capital.
- **Population**: population of the city.

For this dataset we dropped the rows where the country was not in the world countries dataset. We also modify the Country column to CountryID to match the countries table.

1.4 Modification on the World countries dataset

The world countries dataset is a dataset from Kaggle, which contains information on the world countries. The csv file contains the following atributes:

- **Country**: name of the country.
- **Region**: region of the country.
- **Population**: population of the country.
- **Area (sq. mi.)**: area of the country.
- **Pop. Density (per sq. mi.)**: population density of the country.
- **Coastline (coast/area ratio)**: coastline of the country.
- **Net migration**: net migration of the country.
- **Infant mortality (per 1000 births)**: infant mortality of the country.
- **GDP ($ per capita)**: GDP of the country.
- **Literacy (%)**: literacy of the country.
- **Phones (per 1000)**: phones of the country.
- **Arable (%)**: percentage of arable land of the country.
- **Crops (%)**: percentage of crops land of the country.
- **Other (%)**: percentage of other land of the country.
- **Climate**: climate of the country.
- **Birthrate**: birthrate of the country.

- **Deathrate**: deathrate of the country.
- **Agriculture**: some agriculture index of the country.
- **Industry**: some industry index of the country.
- **Service**: some service index of the country.

For this dataset the only modification was to add a new column called CountryID to be the primary key of the table.

2 Modelling the database

ADD THE UML FIGURE HERE.

After cleaning the data we can now organize the tables. We choose to have the following tables with the primary keys in red and the foreign keys underlined:

- **countries**(countryid, country, region, population, area, pop_density, coastline, infant_mortality, phones, arable, crops, other, birthrate, deathrate, agriculture, industry, service)
- **cities**(id, city, city_ascii, lat, lng, countryid, iso2, iso3, admin_name, capital, population)
- **studios**(studioid, name, notable_games, notes, cityid, countryid, year)
- **games**(appid, name, studioid, release_date, required_age, price, dlccount, windows, mac, linux, achievements, estimated_owners, metacritic_score, positive, negative, average_playtime_forever)
- **categories**(name, categoryid)
- **genres**(name, genreid)
- **languages**(name, languageid)
- **game_genres**(appid, genreid)
- **game_categories**(appid, categoryid)
- **game_languages**(appid, languageid)