Members
Michael Bradberry
Tanner Lievois
Tyler Ward

Team 4: ETL Project Final Report

"This project was challenging, and I'm ready for a nap."

- Tyler Ward, Michael Bradberry and Tanner Lievois

Data Analytics Students

Extract

The original dataset from Kaggle was downloaded as a csv file and was fairly comprehensive. What it was missing however was any mention of review data, and for a person who would care to analyze the sales data and compare it to reviews, the team felt that it would be convenient.

The team was able to web scrape Metacritic for game reviews. The additional data came from the below link with the URL being concatenated and composed of the platform and game name (i.e. URL/platform/game_name)

- https://www.metacritic.com/game

1) https://www.kaggle.com/gregorut/videogamesales
2) Dataset provided by the file "Review_Scrape.ipynb" that pulls Metacritic reviews for all of the game/platform combinations in the video game sales data.

Transform

Data was downloaded in first normal form and the transformation required dissection of VG sales data into second normal form. Data cleaning was significant and rows with duplicate or null values were dropped. Sales data was dropped for those platforms weren't available on Metacritic.

A total of 7 tables were created with a combination of one-to-many and one-to-one relationships depending upon the data requirements. Of the tables created, one of the most significant was the title-platform junction table that allows the cross-reference of data.

Load

Because of the quantity of one-to-many relationships, it seemed especially valuable to switch directions of thought and begin designing a relational database instead of the non-relational as indicated by the proposal.

While loading the data, the constraints allowed us to reconsider the quality of the data and revisit the transform phase of the project to rework the dataframes.

Pushing the data to the database was done via SQLAlchemy's connection engine and required login and password via a separate .py file that was imported. This allowed the team to avoid exporting and importing a variety of different csv files into PostgreSQL.

RICE