**ETL Project**

Awe Ryan, Daniel Chavez

horizontal line

# Placeholder image

# ETL can be fun!

We started with the following data sets:

<https://www.kaggle.com/gregorut/videogamesales>

<https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings>

The data sets take a look at video game sales of games exceeding 100k copies worldwide. The data sets include rank, name, platform, year, genre, publish, NA Sales, EU Sales, JP Sales, other sales, Global sales, critic rating, and user rating. For this project we are analyzing video game sales by console, rating, and sales. We hypothesized that this will be a relational data set as not all games are which are rated highly sell the best in our opinion.

We took each CSV file and imported them to pgAdmin. We combined each data set in order to make a more complete CSV file by joining each column's critic scores and user scores to give us a bigger picture of which platforms sold best and had the best rating from this data set. Once in SQL we created tables for sales and rank.

We will combine both data sets into SQL to make our tables one of each corresponding to sales and scores. We made the primary key by the console. We then filtered the sales information by console. We then extracted the data to a jupyter notebook where we joined the tables from our data sets using pandas data frames to sort values. When cleaning the data in pandas we filtered out the NaN values from the sales and critic/user scores in order to give us a more complete sample size. We then performed a query on our data in pgAdmin.

Analyzing the data tables showed us there was a relational correlation between scores and sales for each video game. This went against our theory that the data was nonrelational and showed that highly rated games both by critics and users alike will be the best selling across the charts. The top-selling game to our surprise was Wii sports by Nintendo with a user score of 8/10, a critic score of 76 /100, and 82.53 million units sold globally.