CWRU Data Analytics Group Project: ETL

Blue Team: Nancy Dziak, Marcus Eli Pames, Aaron Ronay

Technical Documentation

October 23, 2018

Data Cleanup & Analysis

Once you have identified your datasets, perform ETL on the data.

Make sure to plan and document the following:

**The sources of data that you will extract from.**

CSV files:

Dataset 1: <https://www.kaggle.com/gregorut/videogamesales/downloads/videogamesales.zip/2>

Dataset 2: <https://www.kaggle.com/egrinstein/20-years-of-games/downloads/20-years-of-games.zip/2>

**The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).**

**Cleaning:**

Imported dependencies and read csv files into pandas.

Created data frames for each CSV file, and selecting only N64 titles for North America, using the following columns:

Sales csv: 'Name', 'Platform', 'Year', 'Genre', 'Publisher', 'NA\_Sales'

Reviews csv: ‘title', 'score', 'score\_phrase', 'editors\_choice'

Rename n64 reviews column from "title" to "name"

Reset index for both data frames.

Import SQL connection dependencies and establish parameters for SQL connection to amazon server.

Convert n64 reviews DF to SQL table, and then upload table to database using SQL connection, non-indexed, because join will be later performed based on "name"

Convert n64 sales DF to SQL table, and then upload table to database using SQL connection, non-indexed, because join will be later performed based on "name"

**The type of final production database to load the data into (relational or non-relational).**

Relational (MySQL) using SQL Alchemy.

**The final tables or collections that will be used in the production database.**

N64 Sales and N64 Reviews, then perform an inner join on the “Name” column to join the two tables.

You will be required to submit a final technical report with the above information and steps required to reproduce your ETL process.