Report of ETL project

Alejandro Velez and Gary Castillo

Extract

We went to the website <https://overwatchleague.com/en-us/statslab> and downloaded all the csv files from there. We extracted the data from the zipped files, then we stored them in our resources folder. We then loaded all our csv’s into our pandas’ data frames. (In the resources you’ll notice two of the data sets are zipped, that’s because they were too large to load into the Github project).

Transform

So right away once we loaded our csv files into a panda’s data frame, we removed the map type, tournament title from the 2021-2020 files, and map type, stage from the 2019-2018 files. We then, from every file, removed any row that had a value not equal to “All Heroes’ in the hero column.

After removing the columns we didn’t want, we then combined our the numerous data frames we build to create 1 large data frame. We then changed the names of multiple columns so that our column name would match each other.

We had 400,000 plus data points with a lot of repeating values. In order to drastically reduce the number of rows in the data frames, we translated the All stats column and its row values into new columns based on the unique values in the All Stats column. Afterwards we had a lot of NaN Values. However, in this dataset, all the NaN values were in slots designated for numerical values, so in the pivot table function, we told it to fill all NaN values to 0.

Afterwards, we needed to alter the new columns values to easily fit into the sql database. We first removed all the spaces and ‘-‘s in the columns and used the lower() function to have them more easily fit the sql database format.

We did that separately for the two datasets we were working with. After our work we then combined the two large data frames to create a large singular data frame and set a numerical primary key.

Load

With all our data transforming done, we then loaded the data frame into a postgres sql database which we then ran numerous queries on to test its effectivenss