Tyler Turk

Kush Patel

ETL Project

For our ETL project we decided to use Kaggle to extract NBA data sets. We found 2017 NBA stats data which displayed the stats (point percentage, minutes played, age, etc.) of NBA players in the 2017 season which was in the form of a CSV file. We also used NBA 2017 Salary data which had each NBA players salary in the form of a JSON file. To transform the data, first we cleaned both data sets while still in CSV and JSON form. We removed duplicate columns and symbols such as percentage signs and spelled it out in order to pass the data over to a SQL database. Once the data was cleaned and able to be passed over to a SQL database, we loaded both CSV and JSON files into a pandas data frame. With the data in a pandas data frame we were able to format the data in order to make it look presentable. Once the cleaning process was complete, we joined the CSV and JSON file on the player name. The final tables that will be loaded into the data base will be a NBA player’s salary, NBA player’s stats table as well as a table with the combined data. We will be using POSTGRES to store our data due to the relationship between the two datasets.