|  |
| --- |
| Gsummit  Technical Report |
| **ETL Project**  **Group 5**   * Marshal * Badou * Pat * Matthew * Yosief * Ray |

Introduction

As a part of Gsummit, you will find that our values are embedded in our work ethic. Today we release our latest ETL, where we Extract, Transform, and Load information based on the correlation between NBA draft picks and the athletes that have accomplished making the all-star team! Today is very proud day to be unveiling our latest and greatest finds. Stay tuned as we organize and dive deeper into explaining the data further.

Project Scope

Gsummit’s aim is to present our findings and technical expertise on ETL to the public. The ETL process will be carried out by cleaning, joining, filtering, and aggregating data found in our csv files obtained from Kaggle. Relational databases such as pgAdmin 4 and MYSQL will be used to create, update, and manifest databases/ tables.

Stage 1: Extract

Our initial step is to do vital research, by using credible websites to bring in datasets that reflect the overall direction of our project. After long deliberation, Gsummit has chosen to search for data based on the Stats on players, teams, and coaches in men's pro basketball leagues, 1937-2012. Kaggle presented numerous datasets that proved useful to our underlying narrative (correlation between NBA draft picks and the athletes that have accomplished making the all-star team). Three csv files have been used to carry out this step in our ETL process (basketball\_coaches.csv, basketball\_awards\_coaches.csv, basketball\_player\_allstar.csv). These three csv files were brought in though Jupyter Notebook using the pd.read\_csv command.

|  |
| --- |
|  |

Stage 2: Transform

The second step in our process after bringing in our CSV files is beginning to transform our datasets so that they are presentable. The overall goal in this step is start to set precedence for our data, so that it seamlessly can be brought into our relational databases and manipulated as needed. We begin by bringing together or joining our three csv files; using the command df.join = pd.merge command. After our csv files are joined and merged, we start to rename columns so that they are more concise and understandable; using the command df\_join.rename(columns={}) . We use the df\_join3.drop(columns={}) command to drop and columns not needed. We use df\_join2.fillna(0) which converts any null value to 0. Lastly, we use the df.to.csv() command to write our pandas Dataframe to an output csv file.

Stage 3: Load

The third and final step in the ETL process is Load. Load is as straightforward as it sounds which why we start by opening our relational database. In this case PostgresSQL, where we start by establishing a database by using CREATE basketballdb;. We review our databases by using the SHOW databases; command. We can choose to use a particular database by using the USE basketballdb; command. Next, we can start creating our tables in our database by using the CREATE draft (column1 datatype,  
column2 datatype, column3 datatype,...), and then navigating to the schema -> basketballdb -> Tables -> draft; where you will right-hand click and selecting Table Data Import wizard. After following these steps our data will be brought in and populate our table as long as our columns, datatype, and sizes are correctly labeled.

Findings:

Once our RDMBS (Relational Database) has executed each SQL statements; we are presented table results. After reviewing our data, we find that there is a solid relationship between NBA players drafted and players that have made the All-Star Team; showing truly gifted players that have been recognized as a premier star in the league like Kareem Abdul Jabbar that has been selected over 18 times! Another finding derived from our dataset is that of NBA players drafted that did not make the All-star Team; Promising college athletes like Kwame Brown who was drafted #1 in 2001 that did not quite reach the level of play as their peers. The last correlation we wanted to observe

statistically which draft pick (grouping draft rank) were more prevalent to become all-stars; astounding as it may seem we found that the number one draft pick in each class accounted for the largest grouping amongst other picks (second place went to the 12th pick overall).

Summary

Overall, Gsummit hopes that you have been able to walk away with a step-by-step understanding of the utilization of ETL. When a user Extracts, the user focuses on the information they may be obtaining and ensuring credible sources to use in this case. The user later Transforms the data, by analyzing the key ways to arrange the data so that it offers the best and most optimal output possible before loading. At this point, the data is finally ready to Load, Our RDBMS are equipped to bring in our dataset into preset tables (user-defined) in our queries were we are able to carry out various SQL statements that enable the user to make a strong correlation between NBA draft picks and the athletes that have accomplished making the all-star team!