**ETL Project by:**

Emre Yener, Dan Smart, Abiel Ogbe

**ATP World Tour Tennis Data**

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

We found two datasets from Datahub Sports Data. We extracted both datasets as csv files. The datasets includes all matches, tournaments and stats for all matches from 1991 through 2017. A total of over 180 thousand rows and 77 columns of data combined.

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

The first thing we did was parse through the data in our jupyter notebook using pandas functions. We extracted the columns with key data stats we felt were most useful for our analysis. We found that both datasets included the ‘match id’ which we figured would be a good way to merge the data frames into one. As we continued to analyze the data, we thought it would be useful to create new columns to calculate percentages on key statistics. For example, we calculated the ‘break point’ percentages for both winners and losers in each match. We then combined the new column of data to show a summary column.

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

Our data was well structured so we decided to use a relational database. PostgreSql proved to be very useful to present the final database. We loaded the our cleaned data from jupyter notebook into PostgreSQL using the ‘create\_engine’ function from the sqlalchemy library.

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

We joined the two datasets on the column ‘match\_id’. We then created a view ‘tennis\_analysis’ to present the key statistics for our analysis. We were able to pull the data back into our jupyter notebook using the pandas function ‘pd.read\_sql\_query’.