**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 ddatasets as csv files. The data sets includes matches and stats for all matches from 1991 through 2016. A total X amount of rows and X amount 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 culmns with key data stats we felt was most useful for our analysis. We found the both datasets included the ‘match id’ which we figured would be a good way merge the dataframes into one. As we continued to analyse the data we thought it would be useful to create new columns to calculate percentages on key statistics. For example we calculated the %%%%% for X .

**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 sql\_alchemy library.

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

We joined the two datasets on ‘match\_id’. We then created a view to present the key statistics for our analysis.