ETL Project

NOVEMBER 16

Doug Watola
Tushar Makhija
Krishna Chawla

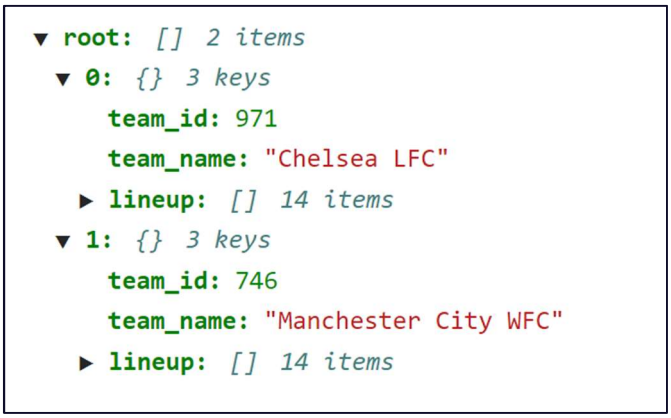
Introduction

What is StatsBomb and why we chose their data for our ETL Project?

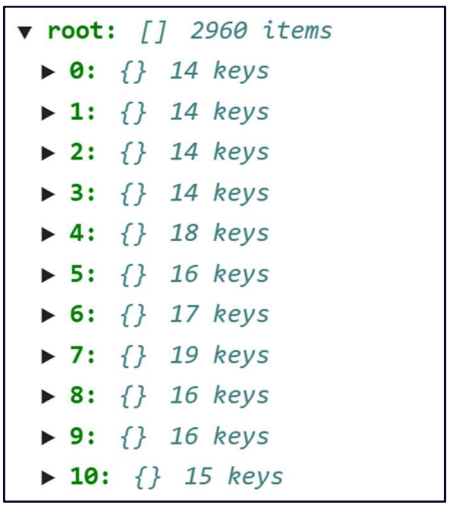
- StatsBomb is a soccer data collection company, that also provides consultations to major soccer clubs around the world. Among other things, soccer clubs find this company helpful in scouting players and manager. StatsBomb data is extensive, but not free, and usually available only to businesses. Fortunately for us, they put out some free data on their GitHub . This gives a glimpse to their data and, enthusiastic people like us can practice on it
- Our vision with this ETL project is to create a meaningful dataset from a player scouting point of view. Therefore, the end goal is to create data on each player and measure things like – pass rate (successful passes / attempted passes), conversion rate (goals scored / shots taken) etc.
- In interest of time, we only were able to demonstrate pass rate, but the process can be replicated for similar or more complex player stats

Step 1: Extract

- Created a clone of this repo: https://github.com/statsbomb/open-data
- The data is in 3 subfolders named:
    - Events
    - Lineups
    - Matches
- Lineups has 637 .JSON files. Each file represents one soccer game. These files have the list of players that made an appearance for each team in the game. Lineup JSONs have 2 children. One for each team. See sample screenshot below



- Events has 637 .JSON files also, which has a 1-1 relationship with lineups files. These files are the core of the data and have details about everything that happened in the game. Each pass made, shot taken, goal etc. Below is a screenshot of a sample event JSON. On average event JSONs have ~ 3,000 children. 2960 for the for one below.



- Formal documentation is on their GitHub: https://github.com/statsbomb/open-data/tree/master/doc
- In the interest of time we did not process the matches files and the related subfolder structure
- The repo we saved all our work on is this: https://github.com/tushmakster/stats_bomb_etl
- Out of the 637 .JSON files we only used 30 to not increase processing time. They are saved on our GitHub repo.

Step 2: Transformation

- All the work done is in a notebook here: **Jupyter Notebook Link**
- Libraries we used were pandas, os, json, and sqlalchemy
- From lineups, we created a unique list of players who had any appearances
- From events, we got passing and shooting data, for each player, for each game
- We want the transformation to be more in depth and more robust but in interest of time we were not able to

Step 3: Loading

- In the last section of the Jupyter notebook reference above, we export the data into a postgres SQL table
- Our SQL code is saved here: Query . Note that I redacted my username/password so be careful if you are trying to recreate everything.
- The vision from there is to have a master table, and multiple SQL views for various kinds of summaries.
- One schema we had in mind was to have 2 SQL views – one for player-based stats, the other for team stats

- Below is a screenshot of one of our final view

Data Output | Explain | Messages | Notifications

	player_name text	passrate numeric
1	Abbie McManus	0.78
2	Abby Dahlkemper	0.79
3	Abby Erceg	0.83
4	Abby Smith	0.62
5	Adriana Leon	0.74
6	Adrianna Franch	0.63
7	Alanna Kennedy	0.74
8	Alex Morgan	0.66
9	Ali Krieger	0.72
10	Allie Long	0.84
11	Allysha Chapman	0.63