Rutgers ETL project – MLB Player Analysis for Potential Online Memorabilia and Apparel Store

Summary

We provide a method that enables potential MLB online store owners for memorabilia and apparel to see what players/teams to have in inventory. We used baseball-reference.com for MLB player information and EBay.com for listings of memorabilia and apparel. The results from these websites were loaded into a MySQL database for analysis.

Pipeline Running

1. Run Player\_Cleaner.py

Narrative/Motivation

We wanted to start an online baseball memorabilia/apparel store starting with the most current players and their teams, but only those that had a significant number of items that were on sale at other online retailers. We didn’t want to carry an inventory of items that wouldn’t sell much.

Final Schema / Data Model / How to use the data

Data Sources

We used baseball-reference.com for the current player and statistical information. baseball-reference.com is recognized as a leader in baseball player statistics gathering and archiving. The assumption of this data is that it is updated on a daily basis. We would only run this process once per week, since significant statistical changes don’t happen daily. EBay.com was used to get the number items that are for sale at any given point. EBay was chosen since there are numerous sellers that we can pull from. Additionally, we were able to scrape the data we needed.

Transformation Step

|  |  |
| --- | --- |
| Program | Step |
| Player cleaner.ipynb | Split out name from name id |
| Player cleaner.ipynb | Removed extraneous characters |
| Player cleaner.ipynb | Added datetimestamp to hitters and pitchers.csv |
| Player cleaner.ipynb | Dropped NaN |
| Player cleaner.ipynb | Removed dupes |
| Player cleaner.ipynb | Selected columns for output .csv files |
|  |  |