Project 2: Extract, Transform, Load
Christina Thip, Charles Glover, Coleman Craig, Lauren Tureaud
Technical Report
11/2/2019

Extract

1. Navigate to kaggle.com, under Datasets tab, search “NFL”
2. Under public datasets, find “NFL and scoring data” uploaded by username spreadspoke
3. Under Data/ Data Sources select each file, 3 total, as follows:
    a. nfl_stadiums.csv
    b. nfl_teams.csv
    c. Spreadspoke_scores.csv
4. For each selected file, underneath the Data table, find file snapshot and select Download

Transform

5. Using Jupyter Notebook, extract data from 3 CSV files in dataframe(s)
6. Rename columns to simplify naming conventions; drop unwanted columns
7. Join (default join) Spreadspoke_scores.csv with nfl_stadiums.csv on column “stadium”
8. Join (default join) newly joined file, named “c_file” with nfl_teams.csv on column “team”
9. Export newly joined masterfile “nfl_file” to CSV
10. You may use Postgres to build your database. Set delimiter to Tab, with commas present in data, lest Postgres misinterpret columns when building database.
11. You may also export data to SQL via Pandas and SQLAlchemy. If using Pandas and SQLAlchemy, commas present in data are a nonissue.

Load

12. Create new Jupyter Notebook; import Pandas and SQLAlchemy
13. Import CSV “nfl_file”
14. Show CSV to ensure that file is found
15. Create connection string using username / password
16. Connect to engine
17. Create table using Pandas
18. Load dataframe to SQL

- Jupyter Notebook is preferred as a method of cleaning data, to easily view progress as you clean and combine data

- Final collection should represent all data deemed relevant from initial assortment of data across multiple CSV files

- Pandas and SQLAlchemy is preferred as a method to load collection for ease and simplicity, such as quick iterative tests. Manually building tables and writing columns in Postgres is preferred where exercising greater control over data structure is preferred beyond Jupyter Notebook