NHL players in entertainment media

The idea behind this project was to combine a dataset of current and former NHL players with a movie dataset, which allowed us to look for their appearance in movies and TV shows. This created a database capable of answering the questions:

1. *How many players has appeared in television?*
   73, including Jaromir Jagr, Don Cherry and Niklas Lidström

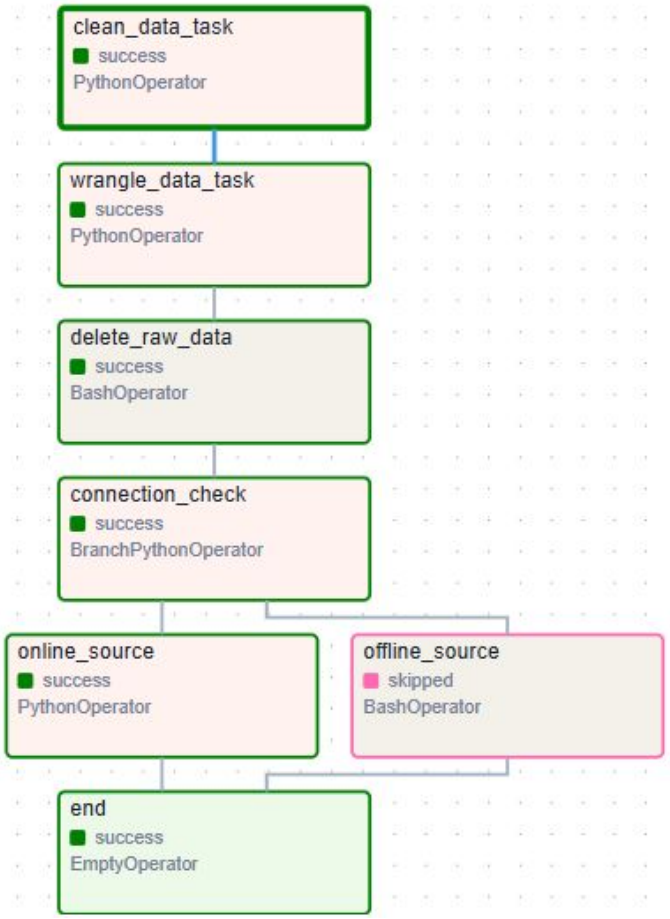
2. *Which year contained the most appearances of NHL players in movies?*
   2010, where 40 players made an appearance.

3. *Which players has made at least two appearances in both tv shows and movies?*
   19, including Wayne Gretzky, Sean Avery and Peter Forsberg

Ingestion

The DAG data_ingestion_dag is used to scrape data about all players who has played a minimum of one game in the NHL from https://www.hockeydb.com. The first step in this DAG is to check if the connection to the website works. If it works, the scraper will loop through all the letters in the alphabet, since the tables are split by the first letter of the last name and then concatenate the resulting data and save it to a csv. If the connection doesn't work, the DAG will instead use a small downloaded subset of the data from the website. When the online or offline scraper has succeeded, the staging DAG will be triggered to start.

Staging



The DAG data_staging_dag then process the data from the ingestion. In the first step, the data is cleaned and in the second it’s transformed to the correct format. The resulting data is saved to a new csv file and the csv file containing the raw data from the ingestion zone is deleted.

After this, the connection to the API is tested. If it works, the processed data is used to enrich the dataset. The players birth dates are used to check that it’s not a namesake. If the connection doesn't work, a downloaded subset will be used.

Production

The DAG sql_dag creates and populates the tables based on the data completed in the staging. It starts by instantiating the necessary tables; Players, Movies, TV and lastly Credits, which contains information about what players have played which roles in which movies and tv shows.

After having instantiated the tables, it then populates the players table using the players data along with the entertainment data. Having populated the players table, it then populates the remaining tables using the players table along with the api data.

All the tables are created and stored in an sql database using the star schema. Access to the database is through PostreSQL.