ShawnRussell2019: Data Warehouse Presentation

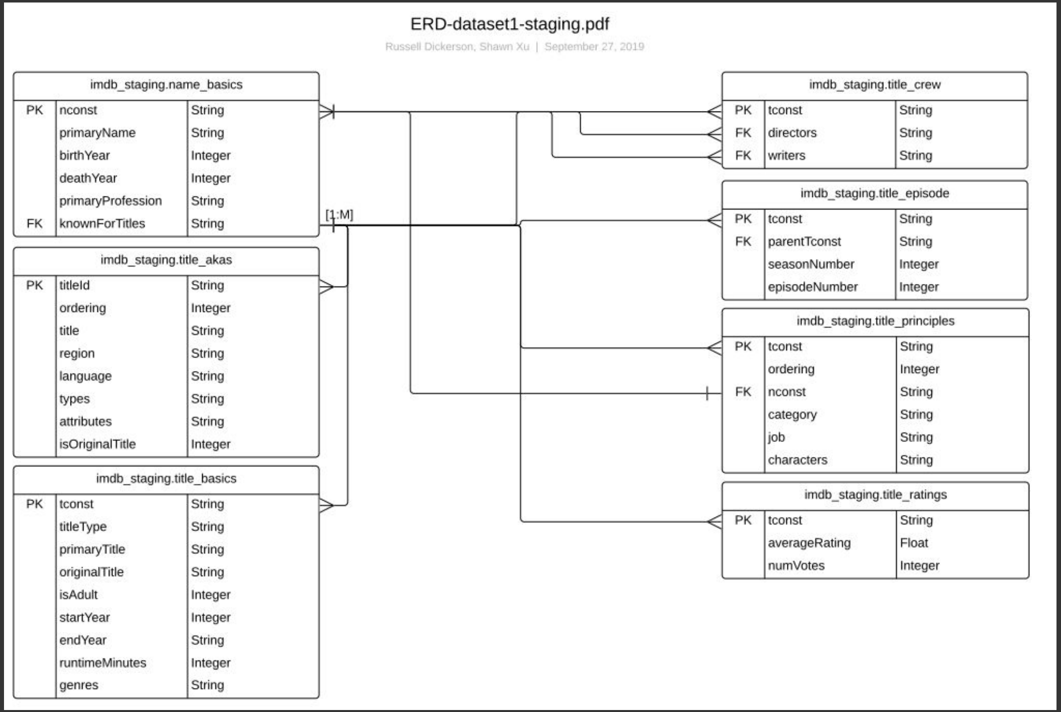
Shawn Xu, Russell Dickerson

Problem/Question to Answer

- By analyzing this dataset we hope to see how the release and attributes of hollywood movies and tv shows is related to/affects the performance of songs on the Billboard charts

Datasets:

- Dataset 1: IMDB
- Dataset 2: Billboard



ERD-dataset1-staging.pdf

Russell Dickerson, Shawn Xu | September 27, 2019

imdb_staging.name_basics

PK	nconst	String
	primaryName	String
	birthYear	Integer
	deathYear	Integer
	primaryProfession	String
FK	knownForTitles	String

imdb_staging.title_akas

PK	titleId	String
	ordering	Integer
	title	String
	region	String
	language	String
	types	String
	attributes	String
	isOriginalTitle	Integer

imdb_staging.title_basics

PK	tconst	String
	titleType	String
	primaryTitle	String
	originalTitle	String
	isAdult	Integer
	startYear	Integer
	endYear	String
	runtimeMinutes	Integer
	genres	String

imdb_staging.title_crew

PK	tconst	String
FK	directors	String
FK	writers	String

imdb_staging.title_episode

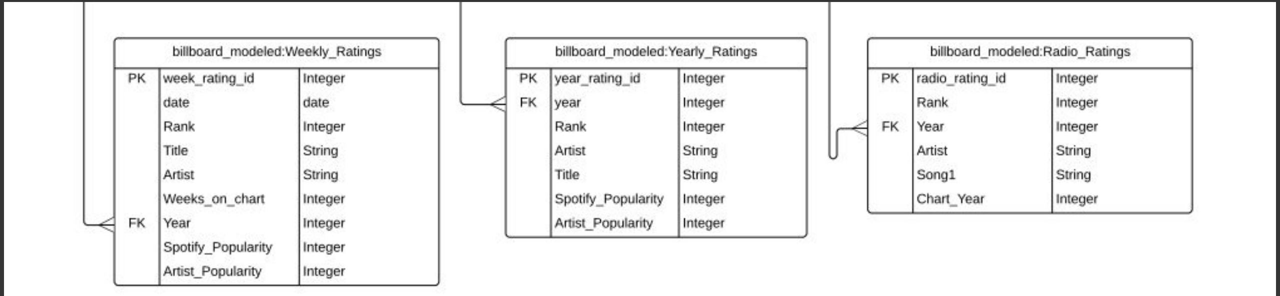
PK	tconst	String
FK	parentTconst	String
	seasonNumber	Integer
	episodeNumber	Integer

imdb_staging.title_principles

PK	tconst	String
	ordering	Integer
FK	nconst	String
	category	String
	job	String
	characters	String

imdb_staging.title_ratings

PK	tconst	String
	averageRating	Float
	numVotes	Integer



billboard_modeled:Weekly_Ratings

PK	week_rating_id	Integer
	date	date
	Rank	Integer
	Title	String
	Artist	String
	Weeks_on_chart	Integer
FK	Year	Integer
	Spotify_Popularity	Integer
	Artist_Popularity	Integer

billboard_modeled:Yearly_Ratings

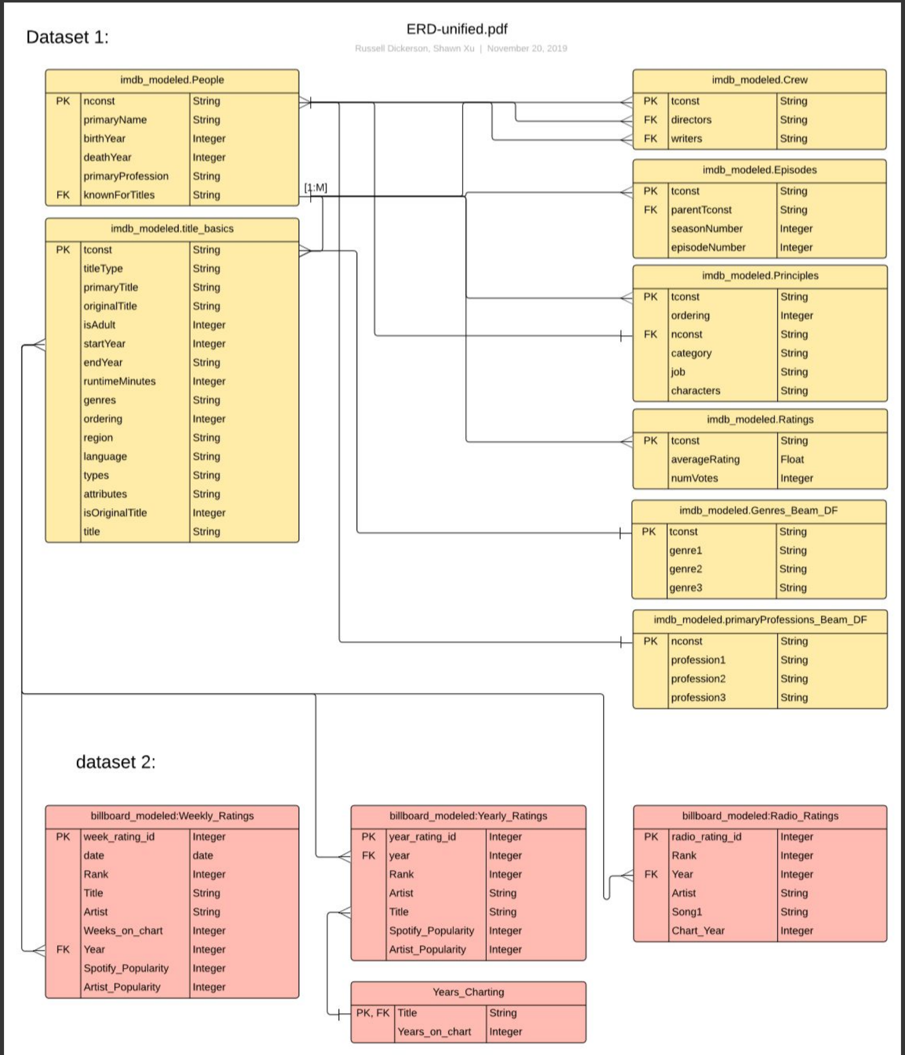
PK	year_rating_id	Integer
FK	year	Integer
	Rank	Integer
	Artist	String
	Title	String
	Spotify_Popularity	Integer
	Artist_Popularity	Integer

billboard_modeled:Radio_Ratings

PK	radio_rating_id	Integer
	Rank	Integer
FK	Year	Integer
	Artist	String
	Song1	String
	Chart_Year	Integer

Modeled Tables

- Dataset 1: Separated “genres” and “professions” columns into individual columns since they had multiple genres and professions in one column
- Dataset 2: Created new entity type “Years_Charting” to relate between years in the 2 datasets



Dataset 1:

ERD-unified.pdf

Russell Dickerson, Shawn Xu | November 20, 2019

imdb_modeled.People

PK	nconst	String
	primaryName	String
	birthYear	Integer
	deathYear	Integer
	primaryProfession	String
FK	knownForTitles	String

imdb_modeled.title_basics

PK	tconst	String
	titleType	String
	primaryTitle	String
	originalTitle	String
	isAdult	Integer
	startYear	Integer
	endYear	String
	runtimeMinutes	Integer
	genres	String
	ordering	Integer
	region	String
	language	String
	types	String
	attributes	String
	isOriginalTitle	Integer
	title	String

imdb_modeled.Crew

PK	tconst	String
FK	directors	String
FK	writers	String

imdb_modeled.Episodes

PK	tconst	String
FK	parentTconst	String
	seasonNumber	Integer
	episodeNumber	Integer

imdb_modeled.Principles

PK	tconst	String
	ordering	Integer
FK	nconst	String
	category	String
	job	String
	characters	String

imdb_modeled.Ratings

PK	tconst	String
	averageRating	Float
	numVotes	Integer

imdb_modeled.Genres_Beam_DF

PK	tconst	String
	genre1	String
	genre2	String
	genre3	String

imdb_modeled.primaryProfessions_Beam_DF

PK	nconst	String
	profession1	String
	profession2	String
	profession3	String

dataset 2:

billboard_modeled:Weekly_Ratings

PK	week_rating_id	Integer
	date	date
	Rank	Integer
	Title	String
	Artist	String
	Weeks_on_chart	Integer
FK	Year	Integer
	Spotify_Popularity	Integer
	Artist_Popularity	Integer

billboard_modeled:Yearly_Ratings

PK	year_rating_id	Integer
FK	year	Integer
	Rank	Integer
	Artist	String
	Title	String
	Spotify_Popularity	Integer
	Artist_Popularity	Integer

billboard_modeled:Radio_Ratings

PK	radio_rating_id	Integer
	Rank	Integer
FK	Year	Integer
	Artist	String
	Song1	String
	Chart_Year	Integer

Years_Charting

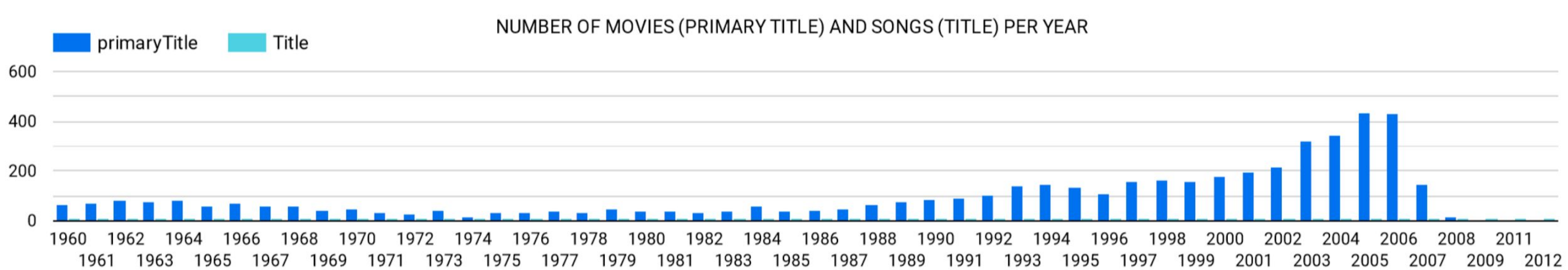
PK, FK	Title	String
	Years_on_chart	Integer

Beam Pipelines

- Fixed IMDB formatting issues using Pardos
  - Fix #1: formatting of genres column
  - Fix #2: formatting of primary professions
- Created new data column using Pardos for dataset consistency in Billboard
  - Fix#3: creating Years_charting column in Yearly_rankings to match corresponding Weeks_charting column in Weekly_rankings

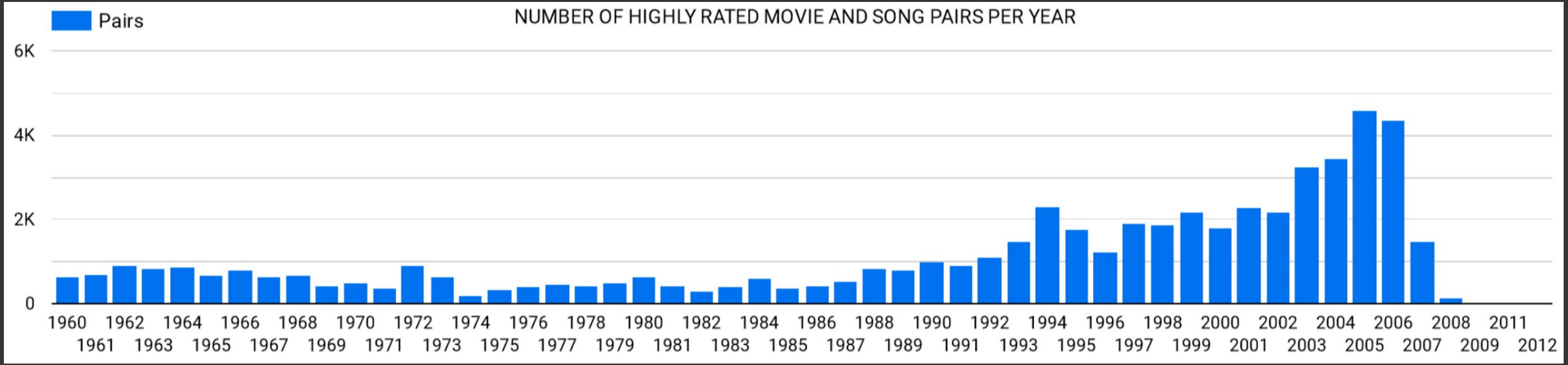
Cross-Dataset Queries:

- Find year and title where the song and movie both have high rankings/ratings

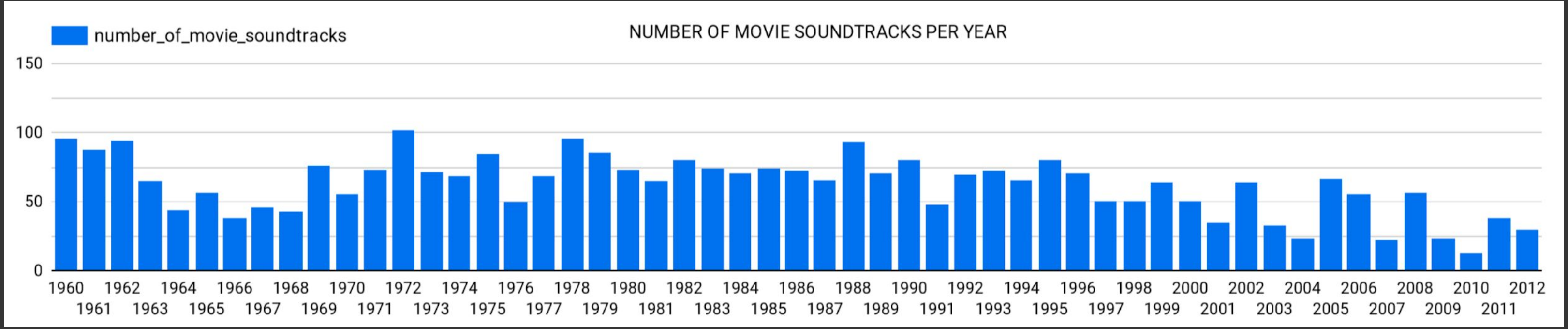


Cross-Dataset Queries:

- Find the number of high rating movies and song release pairs grouped by year

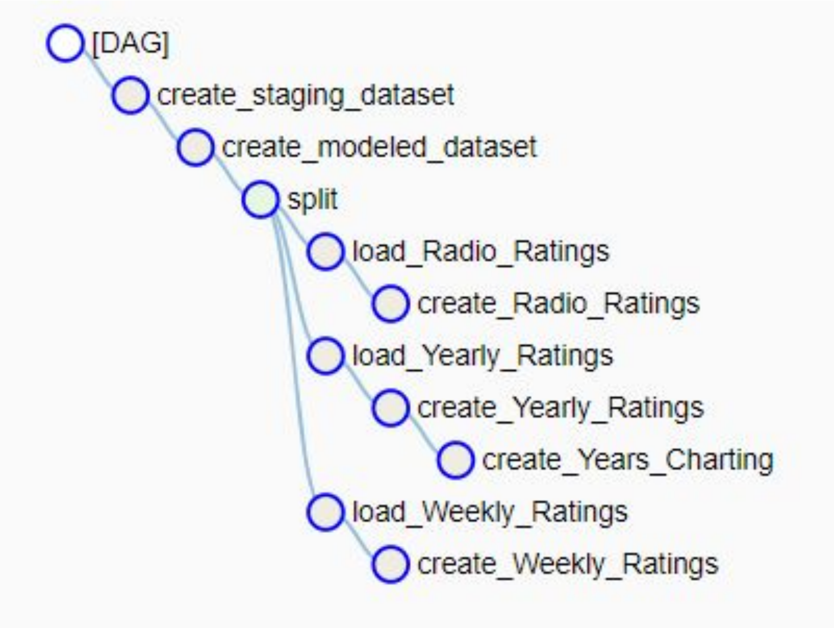


- Find the number of movie soundtracks per year



Airflow DAG

1. Creates staging dataset and modeled dataset
2. Populates staging dataset by loading all csv files at the same time
3. Populates modeled dataset by creating modeled datasets at the same time, except for Years_Charting



Demonstration

- Popularity of movie soundtracks

Future Improvements

- Rankings vs. Ratings
- More Data
  - No. of listeners
  - Song ratings
  - More songs not on charts
  - Regional data on songs