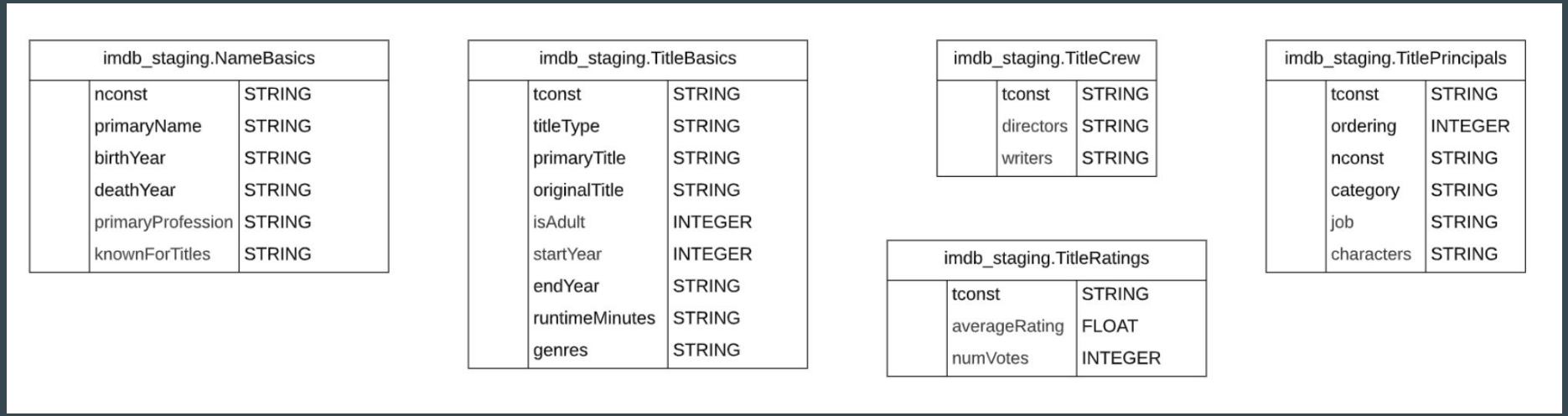
IMDb and Bollywood

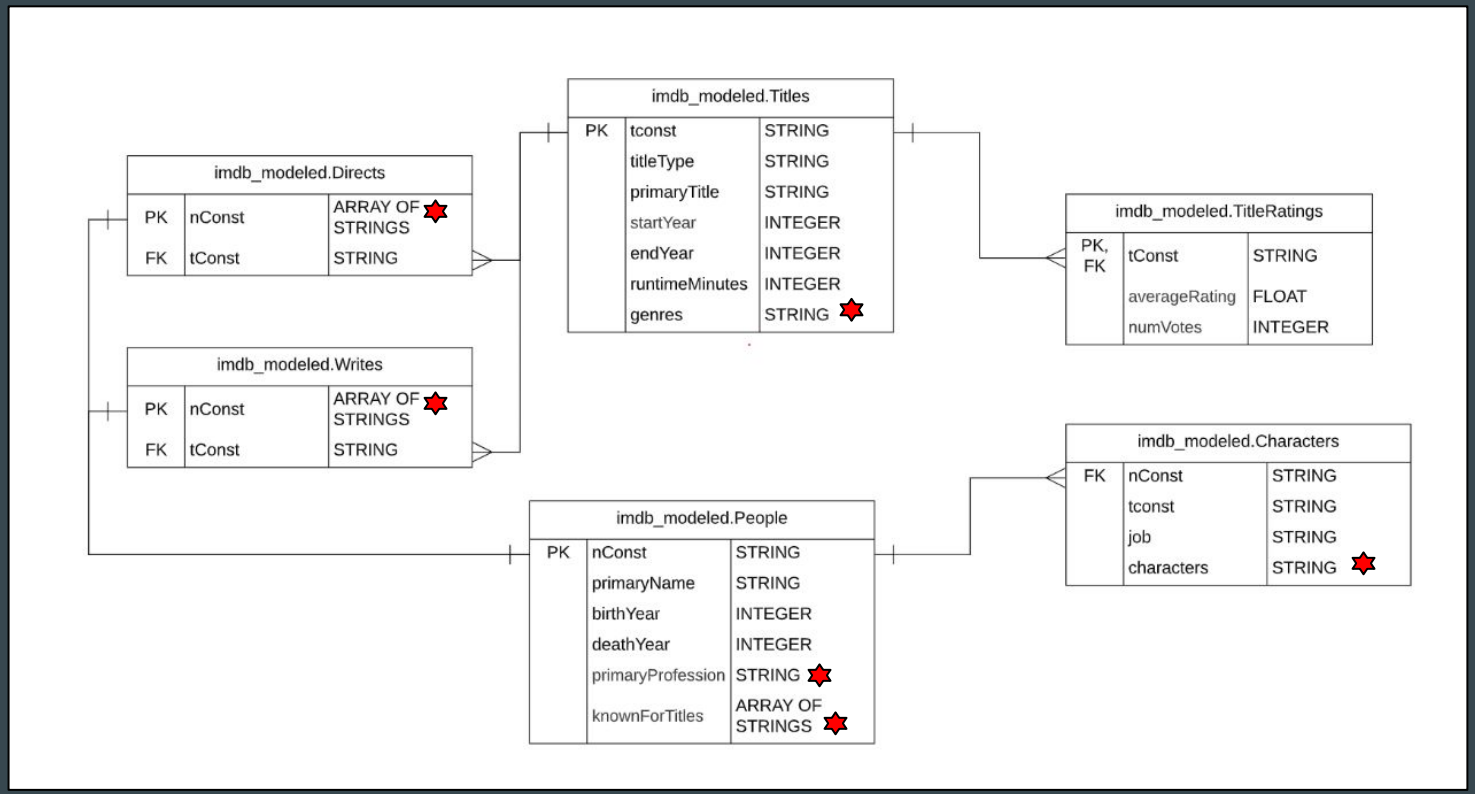
Jacob Helwig and Ian Chang

Raw IMDb Tables



- No primary / foreign keys
- Issues with data types (years, runtime)
- Multiple entities in TitleCrew

SQL Transforms



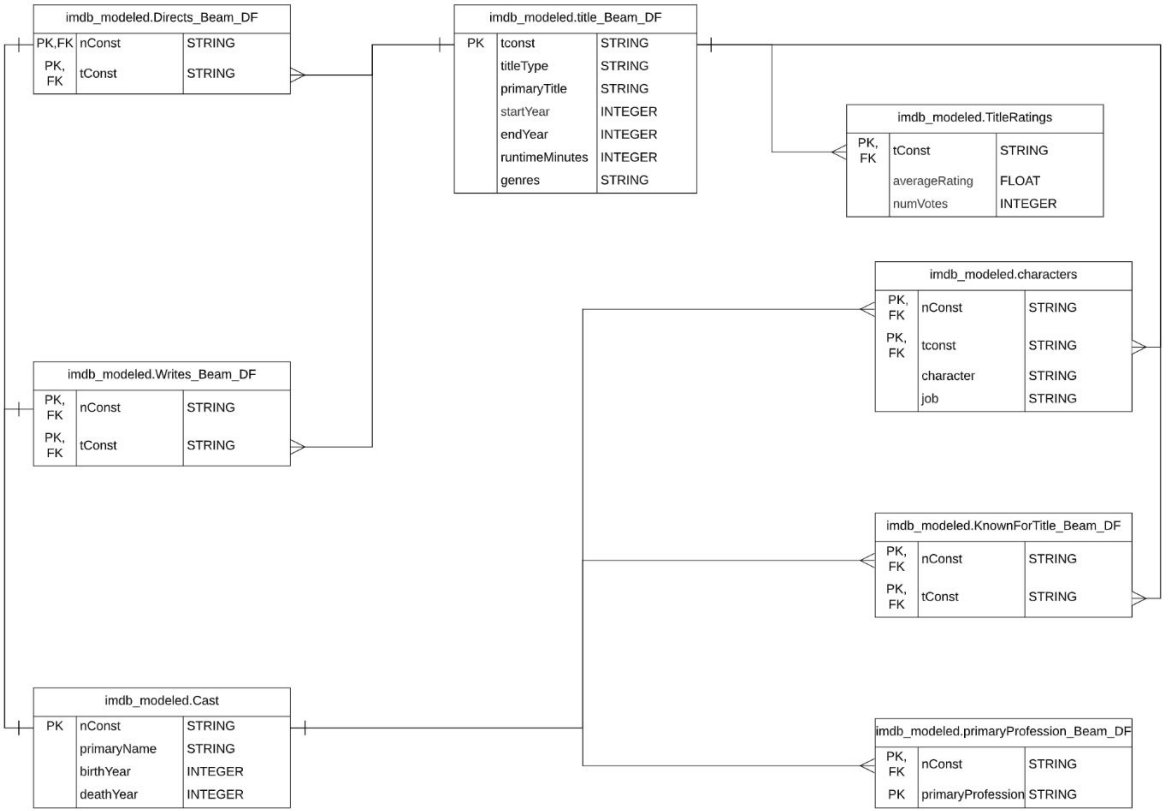
SQL Transforms

- Remove duplicates to establish primary keys
- Break TitleCrew into two tables, only add directors with titles in Titles
- Fix Data Type Issues

Remaining Issues

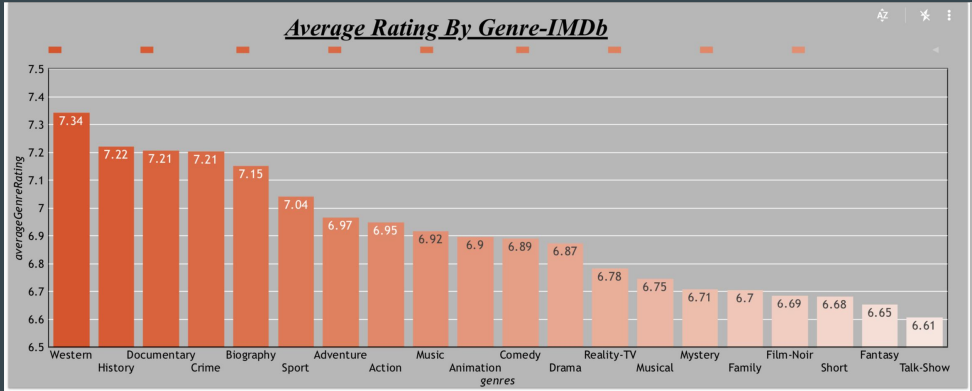
- 1NF violations (6)
- 2NF violations (primaryProfessions, knownForTitles)
- No primary key in Characters table

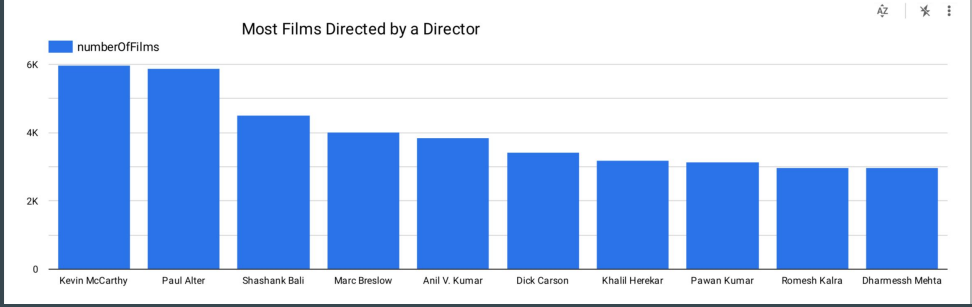
Beam Transforms

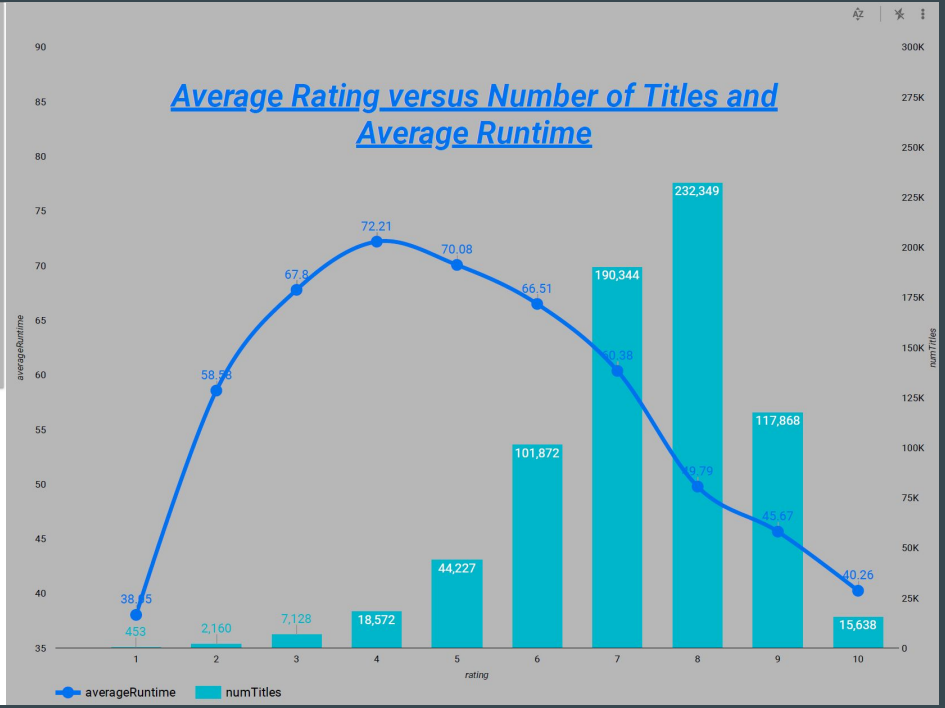


- Resolved 6 1NF issues
- Established Composite PK relationship in Directs and Writes
- Using the SQL query in beam, FK restriction on nConst in Writes and Directs, and on tConst in knownForTitle
- 2NF violations avoided by breaking up People
- Established composite PK in knownForTitle and primaryProfession

Exploratory Data Analysis







Secondary Dataset: Bollywood DB, Raw Data

Bollywood

	Year	Title	Director	Cast	Genre	Release_Month	Release_Date	Highest_Grosser_By_Year_in_crores_
0	1920	Shakuntala	Shree Nath Patankar	unknown	mythology,drama	None	None	None
1	1921	Belgian Emperor's Visit To India	Nitin Bose	unknown	documentary	None	None	None
2	1921	Bilet Pherat a.k.a. England Returned	Dhirendranath Ganguly	dhirendranath,chakraborty,sushilabala,kunjalal...	social,comedy	None	None	None
3	1921	King Gopichand a.k.a. Gopichand	Vishnupant Divekar	mama,koregaonkar,hira,bhatt	legend	None	None	None
4	1922	Bhartrahari a.k.a. King Bhartrahari	S. N. Patankar	tara,koregaonkar,thatte	legend	None	None	None

Actors

	Name	Height_in_cm_
0	Aamir Khan	163
1	Himesh Reshamiya	163
2	Kamal Haasan	165
3	Kunal Khemu	166
4	Shahid Kapoor	167

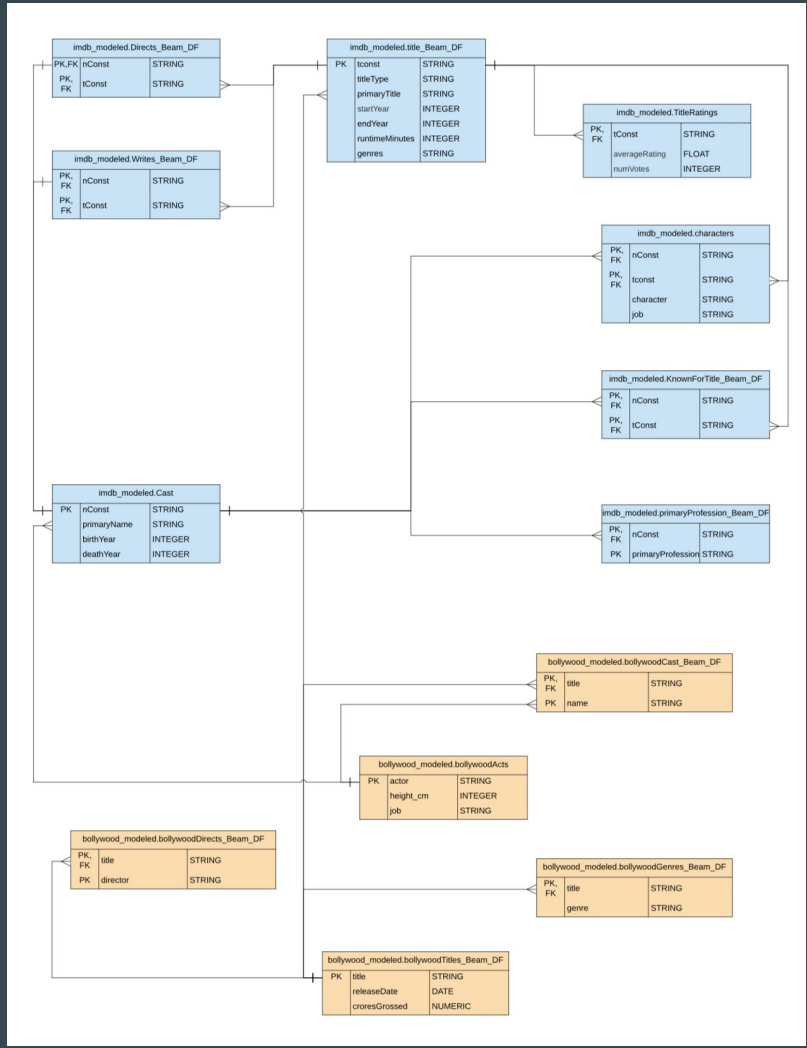
Actresses

	Name	Height_in_cm_	Debut_aslead_role	____
0	Ameesha Patel	152	Kaho Naa... Pyaar Hai	None
1	Soha Ali Khan	152	Rang De Basanti	None
2	Tanuja Samarth	152	Hamari Yaad Aayegi	None
3	Alia Bhatt	155	Student of the Year	None
4	Farida Jalal	155	Laal Patthar	None

- Multiple entities in Bollywood table
- Date in 3 columns; data type issues (crores as String)
- 1NF in cast, genres
- Similar entities in Actors and Actresses; height was a string

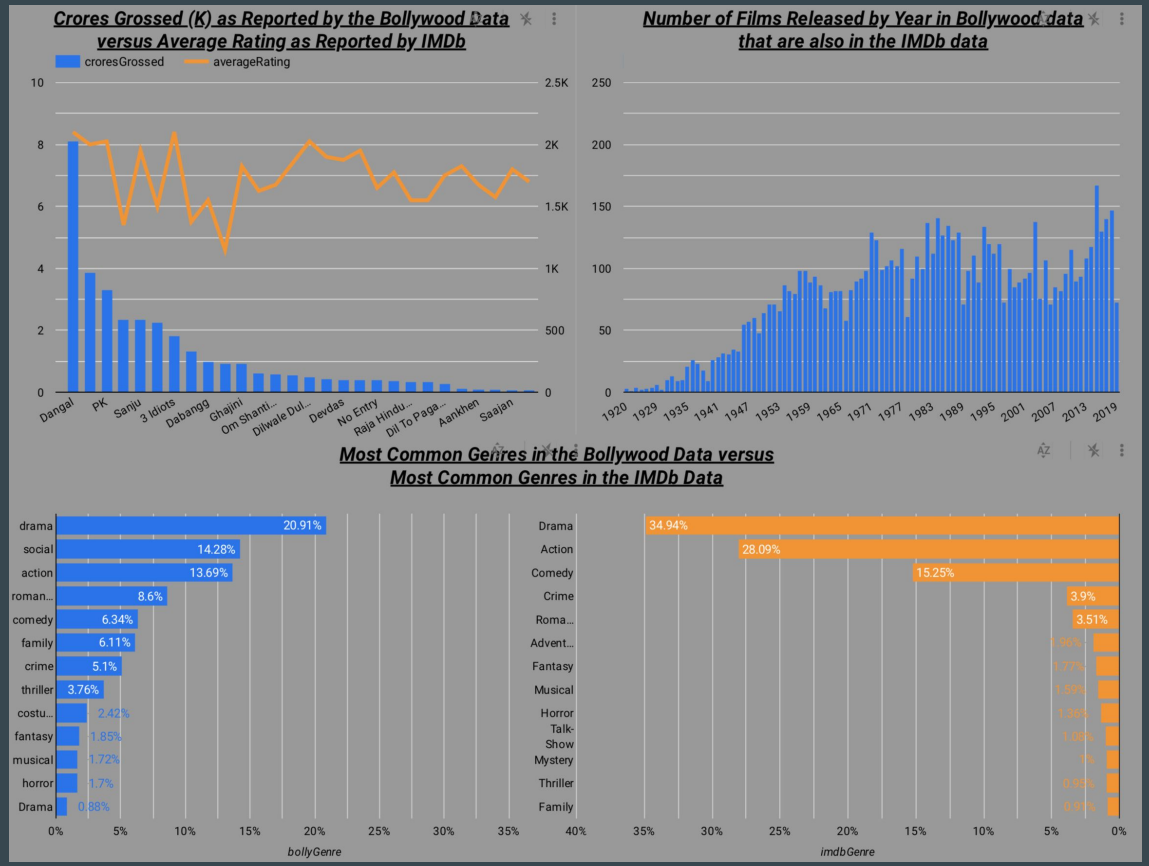
Airflow to Automate Dataflow Transforms

- Break bollywood into Titles, Cast, Directs, and Genre to avoid 2NF issues and 1NF issues
- Combine Actor and Actress
- No unique identifier for movies in bollywood (i.e. tConst)
- Overlap between the data sets



Cross-dataset visualizations

- Look at the croresGrossed reported by the bollywood db versus the average rating reported by the imdb db
- Shared movies between imdb_modeled and bollywood modeled
- Compare counts of genres in each data set



Q&A