DATA ENGINEERING PLATFORMS (MSCA 31012)

Analysis of Movies on Streaming Platforms

Presented by Team Algo:
Han-Yi Lin, Vanshika Tibarewalla, Jacqueline Pezan, Muhammad Ali Ahmad

AGENDA

- Executive Summary
- Research Objective & Business Use Case
- Data Profile
- Data Processing
- Data Modeling and Design
- Analytics and Insights
- Lessons Learned
- What next?

MEET THE TEAM

HANYI LIN

Chief Executive Officer
Data Scientist

VANSHIKA TIBAREWALLA

Executive Vice President
Data Visualization Expert

JACQUELINE PEZAN

Chief Analytics Officer
Data Scientist

MUHAMMAD ALI AHMAD

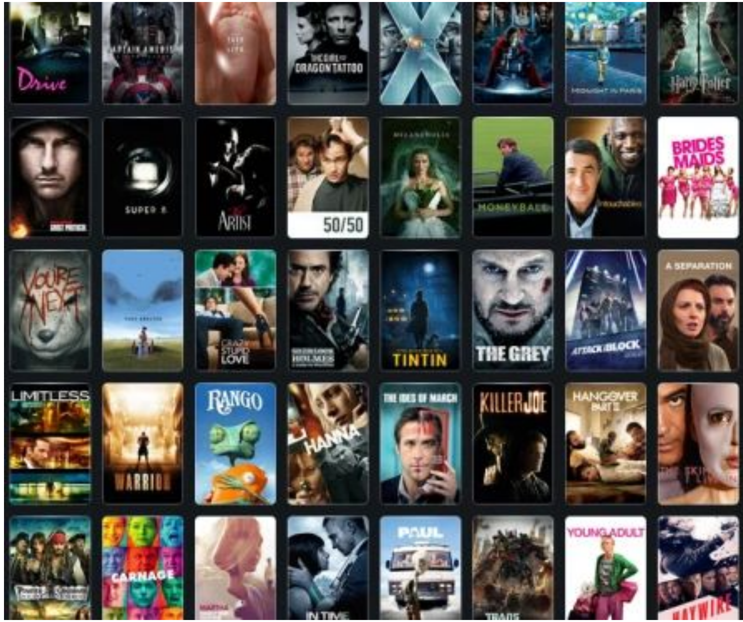
Chief Technical Officer
Data Architect

EXECUTIVE SUMMARY

With thousands of movies available on streaming platforms, users are spoilt for choice. After wasting precious time unable to decide, it all comes down to ratings and reviews. But with multiple ratings, whom does one give priority to?

To address this, we created a master movie database, of movies available on streaming platforms such as Netflix, Amazon Prime, Disney+ and Hulu, combining it with ratings from IMDb, Rotten Tomatoes, Metacritic and others; to solve users problem of plenty.

We developed an interactive dashboard and obtained insights comparing different scoring systems.



RESEARCH OBJECTIVE & BUSINESS USE CASE

- To analyse content available on streaming platforms
- To draw insights from it based on factors such as genre, country, director, year etc
- To create an interactive movie dashboard based on ratings and platforms

- What should I watch today on this streaming platform?
- Which streaming platform can I find this movie on?
- What is the highest rated movie in this category?

DATA COLLECTION & PROFILE

Data Source	Format and Size	Rows/ Cols
Movie Basic Dataset (title, release year, genre)	Structured TSV File 684 MB	1.04 M rows 10 cols
Principal Cast Dataset (director, composer, producer)	Structured TSV File 1.95 GB	1.04 M rows 7 cols
Movies on Streaming Platforms Dataset (Netflix, Prime, Hulu, Disney+)	Structured CSV File 1.18 MB	9516 rows 16 cols
IMDb Rating Dataset	Web Scraping	
Rotten Tomatoes Rating Dataset	Web Scraping	
Metacritic Rating Dataset	Web Scraping	

DATA PROCESSING

DATA IMPLEMENTATION TOOLS

Data Processing	Data Warehouse	Analytics & Visualization	Presentation
python™	MySQL®	python™	P
IMDb API	DG	tableau	Google Slides
X			
OpenRefine			

DATA PROCESSING - IMDbPY

- Use IMDbPY Python package for retrieving data of the IMDb movie database

TOP 250 MOVIES

```
[ ]  top250_mov = ia.get_top250_movies()
     for movie in top250_mov:
         print(movie['title'], movie['rating'], movie['year'])

     The Shawshank Redemption 9.2 1994
     The Godfather 9.1 1972
     The Godfather: Part II 9.0 1974
     The Dark Knight 9.0 2008
     12 Angry Men 8.9 1957
     Schindler's List 8.9 1993
     The Lord of the Rings: The Return of the King 8.9 2003
     Pulp Fiction 8.8 1994
     The Good, the Bad and the Ugly 8.8 1966
     The Lord of the Rings: The Fellowship of the Ring 8.8 2001
     Fight Club 8.8 1999
     Forrest Gump 8.7 1994
```

	rating	title	year
0	9.2	The Shawshank Redemption	1994
1	9.1	The Godfather	1972
2	9.0	The Godfather: Part II	1974
3	9.0	The Dark Knight	2008
4	8.9	12 Angry Men	1957
..	...	...	...
246	8.0	The Princess Bride	1987
247	8.0	Paris, Texas	1984
248	8.0	96	2018
249	8.0	Drishyam 2	2021
250	8.0	Drishyam 2	2021

- Datasets within IMDb database: Top 250 movies top 250 tv, top 100 movies, top 100 tv, bottom 100 movies

DATA PROCESSING - WEB SCRAPING

- Extract rating data by providing a valid IMDb ID
- IMDb, Metacritic, TheMovieDb, RottenTomatoes, TV.com, FilmAffinity

```python
movie_list = []
for i in movie_list2:
    url = 'https://imdb-api.com/en/API/Ratings/k_yqfm5p65/{id}'.format(id=i)
    rating_data = requests.get(url).json()
    movie_list.append(pd.DataFrame(rating_data,index=[0]))
imdb_data = pd.concat(movie_list).reset_index()
imdb_data
```

	index	imDbId	title	fullTitle	type	year	imDb	metacritic	theMovieDb	rottenTomatoes	tV_com	filmAffinity	errorMessage
0	0	tt1302006	The Irishman	The Irishman (2019)	Movie	2019	7.8	94	7.7	95		7.3	
1	0	tt5074352	Dangal	Dangal (2016)	Movie	2016	8.4		8.0			7.4	
2	0	tt11989890	David Attenborough: A Life on Our Planet	David Attenborough: A Life on Our Planet (2020)	Movie	2020	9	72	8.6	95		8.1	
3	0	tt0169102	Lagaan: Once Upon a Time in India	Lagaan: Once Upon a Time in India (2001)	Movie	2001	8.1	84	7.4	95		7.0	
4	0	tt0384766	Rome	Rome (TV Series 2005–2007)	TVSeries	2005	8.7	70	8.2	86	8.8	7.8	

DATA PROCESSING - IMDb CSV FILES

- Collect and clean movie data including country, genres, directors, streaming platforms

	tconst	titleType	primaryTitle	originalTitle	isAdult	startYear	endYear	runtimeMinutes	genres
1.	tt0000001	short	Carmencita	Carmencita	0	1894	\N	1	Documentary,Short
2.	tt0000002	short	Le clown et ses chiens	Le clown et ses chiens	0	1892	\N	5	Animation,Short
3.	tt0000003	short	Pauvre Pierrot	Pauvre Pierrot	0	1892	\N	4	Animation,Comedy,Romance
4.	tt0000004	short	Un bon bock	Un bon bock	0	1892	\N	12	Animation,Short
5.	tt0000005	short	Blacksmith Scene	Blacksmith Scene	0	1893	\N	1	Comedy,Short
6.	tt0000006	short	Chinese Opium Den	Chinese Opium Den	0	1894	\N	1	Short
7.	tt0000007	short	Corbett and Courtney Before the Kinetograph	Corbett and Courtney Before the Kinetograph	0	1894	\N	1	Short,Sport
8.	tt0000008	short	Edison Kinetoscopic Record of a Sneeze	Edison Kinetoscopic Record of a Sneeze	0	1894	\N	1	Documentary,Short
9.	tt0000009	short	Miss Jerry	Miss Jerry	0	1894	\N	40	Romance,Short
10.	tt0000010	short	Leaving the Factory	La sortie de l'usine Lumière à Lyon	0	1895	\N	1	Documentary,Short
11.	tt0000011	short	Akrobatisches Potpourri	Akrobatisches Potpourri	0	1895	\N	1	Documentary,Short
12.	tt0000012	short	The Arrival of a Train	L'arrivée d'un train à La Ciotat	0	1896	\N	1	Documentary,Short

	ID	Title	Year	Age	IMDb	RottenTomatoes	Netflix	Hulu	PrimeVideo	Disney	Type	Directors	Genres	Country
1.	1	The Irishman	2019	18+	7.8/10	98/100	1	0	0	0	0	Martin Scorsese	Biography,Crime,Drama	United States
2.	2	Dangal	2016	7+	8.4/10	97/100	1	0	0	0	0	Nitesh Tiwari	Action,Biography,Drama,Sport	India,United States,United Kingdom,Australia,Kenya,Namib
3.	3	David Attenborough: A Life on Our Planet	2020	7+	9.0/10	95/100	1	0	0	0	0	Alastair Fothergill,Jonathan Hughes,Keith Scholey	Documentary,Biography	United Kingdom
4.	4	Lagaan: Once Upon a Time in India	2001	7+	8.1/10	94/100	1	0	0	0	0	Ashutosh Gowariker	Drama,Musical,Sport	India,United Kingdom
5.	5	Roma	2018	18+	7.7/10	94/100	1	0	0	0	0		Action,Drama,History,Romance,War	United Kingdom,United States
6.	6	To All the Boys I've Loved Before	2018	13+	7.1/10	94/100	1	0	0	0	0	Susan Johnson	Comedy,Drama,Romance	United States
7.	7	The Social	2020	13+	7.6/10	93/100	1	0	0	0	0	Jeff Orlowski	Documentary,Drama	United States

DATA PROCESSING - IMDb CSV FILES

Splitting strings so that we can extract values as separate columns within our data.
Ensuring that we can maintain data integrity and export to CSV to upload to MySQL

```
name_basics_1[['Profession_1', 'Profession_2', 'Profession_3']] = name_basics_1['primaryProfession'].str.split(',', expand=True)

name_basics_1[['Title_1', 'Title_2', 'Title_3', 'Title_4', 'Title_5', 'Title_6']] = name_basics_1['knownForTitles'].str.split(',', expand=True)
```

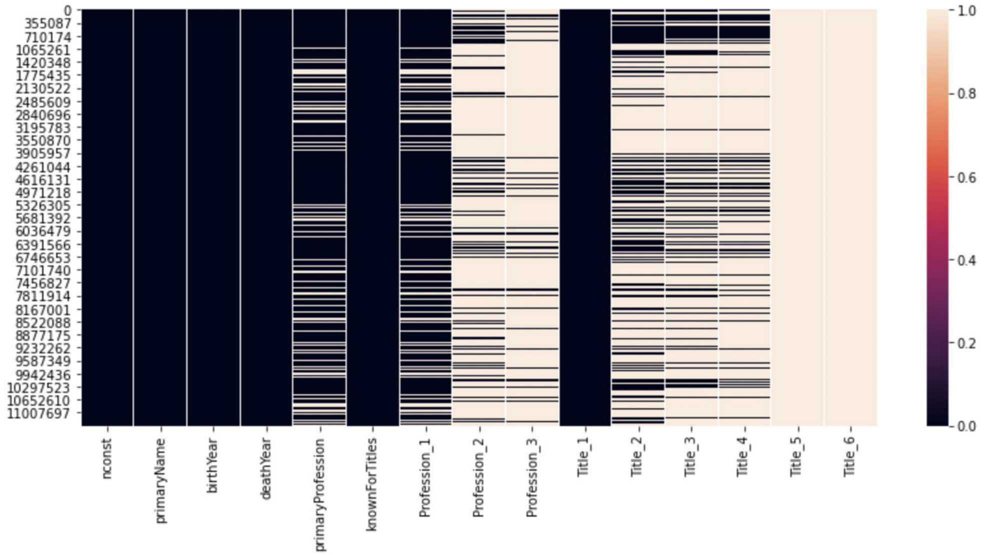
primaryProfession	knownForTitles
soundtrack,actor,miscellaneous	tt0031983,tt0053137,tt0050419,tt0072308
actress,soundtrack	tt0038355,tt0117057,tt0037382,tt0071877
actress,soundtrack,music_department	tt0049189,tt0054452,tt0056404,tt0057345
actor,soundtrack,writer	tt0078723,tt0077975,tt0080455,tt0072562
writer,director,actor	tt0050976,tt0083922,tt0060827,tt0050986

	Profession_1	Profession_2	Profession_3	Title_1	Title_2	Title_3	Title_4	Title_5	Title_6
0	soundtrack	actor	miscellaneous	tt0031983	tt0053137	tt0050419	tt0072308	None	None
1	actress	soundtrack	None	tt0038355	tt0117057	tt0037382	tt0071877	None	None
2	actress	soundtrack	music_department	tt0049189	tt0054452	tt0056404	tt0057345	None	None
3	actor	soundtrack	writer	tt0078723	tt0077975	tt0080455	tt0072562	None	None
4	writer	director	actor	tt0050976	tt0083922	tt0060827	tt0050986	None	None
5	actress	soundtrack	producer	tt0034583	tt0077711	tt0036855	tt0038109	None	None
6	actor	soundtrack	producer	tt0033870	tt0043265	tt0034583	tt0042593	None	None
7	actor	soundtrack	director	tt0078788	tt0068646	tt0070849	tt0047296	None	None
8	actor	soundtrack	producer	tt0087803	tt0057877	tt0059749	tt0061184	None	None
9	actor	soundtrack	director	tt0042041	tt0029870	tt0031867	tt0035575	None	None

DATA CLEANING

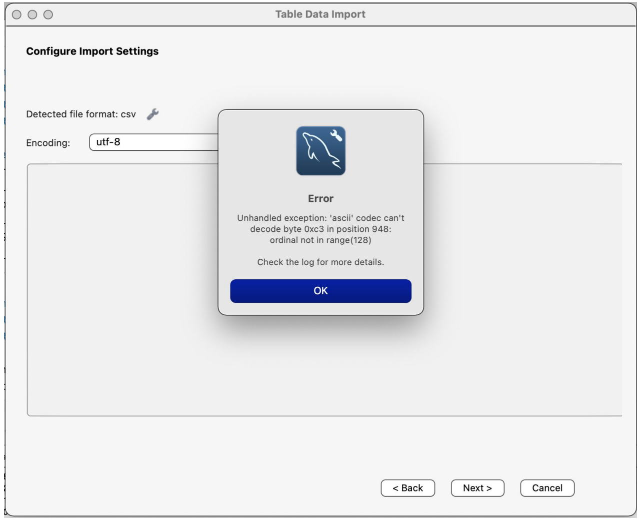
Checking for Null Values within all graphs

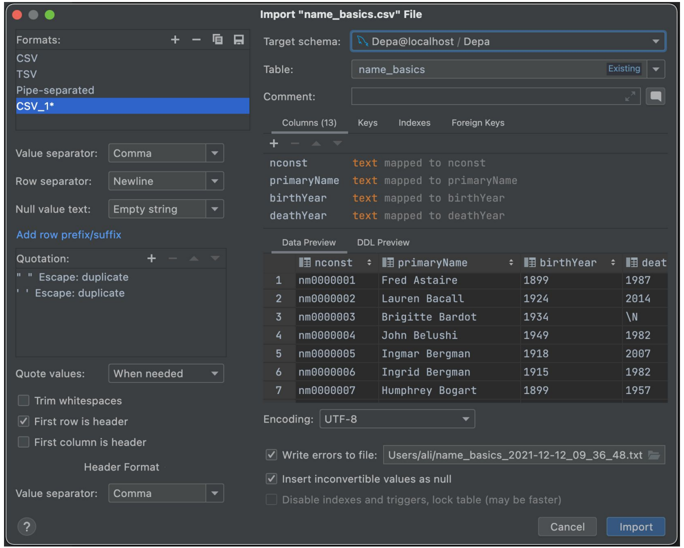
```
plt.figure(figsize=(14,6))
sns.heatmap(name_basics_1.isnull())
plt.show()
```



EXPORTING DATA TO MySQL

Running into errors while using the data import wizard built into MySQL
Fixing this while trying to encode as UTF-8 when exporting from python but running into the same error.
Utilizing datagrip to connect to the relational database, ignore errors and upload our csv data.



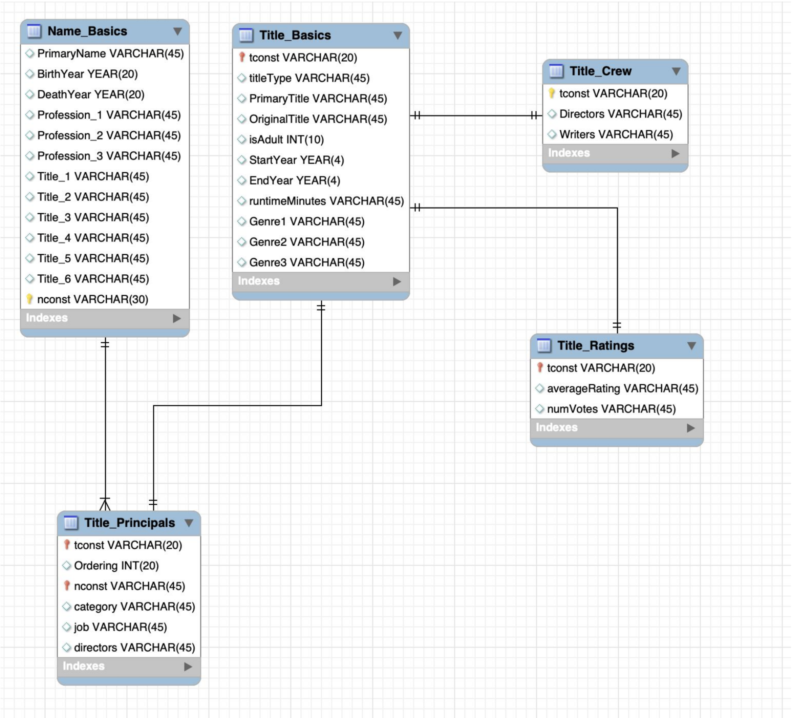


DATA MODELING & DESIGN

DESIGN CONSIDERATIONS

- Identify possible dimensions and related attributes from the IMDb dataset
- Define data type for each attribute (INT for primary key, TIMESTAMP for date)
- Adopt consistent naming conventions (plural table names, snake_case, column alias)
- Use unique identifiers and primary keys
- Ensure the integrity of our joins to make sure data representation remains accurate
- Working with temporary tables database schema to add additional layer of security
- Store final data table in our original schema

Overview of Available Data



Summary of the data available to us
Overview of fields and the relevant data types
How do tables link to each other

DATA PLATFORM

Performing string manipulation to convert our fields to integers and creating indexes. .
Joining on relevant fields to aggregated create tables with all pertinent information.

```
DROP TEMPORARY TABLE IF EXISTS tempdb.`name_basics_0`;
CREATE TEMPORARY TABLE IF NOT EXISTS tempdb.`name_basics_0`
SELECT *, RIGHT(nconst, 7),
RIGHT(title_1, 8) AS title_11, RIGHT(title_2, 8) AS title_22, RIGHT(title_3, 8) AS title_33,
RIGHT(title_4, 8) AS title_44, RIGHT(title_5, 8) AS title_55, RIGHT(title_6, 8) AS title_66
FROM name_basics;

DROP TEMPORARY TABLE IF EXISTS tempdb.`name_basics`;
CREATE TEMPORARY TABLE IF NOT EXISTS tempdb.`name_basics`
SELECT
    nconst, primaryName, birthYear,
    deathYear, Profession_1, Profession_2,
    Profession_3,
    Title_1, Title_2, Title_3, Title_4, Title_5, Title_6,
    REPLACE(Title_11, 't', '9') AS title_01,
    REPLACE(Title_22, 't', '9') AS title_02,
    REPLACE(Title_33, 't', '9') AS title_03,
    REPLACE(Title_44, 't', '9') AS title_04,
    REPLACE(Title_55, 't', '9') AS title_05,
    REPLACE(Title_66, 't', '9') AS title_06
FROM tempdb.`name_basics_0`;
```

```
DROP TEMPORARY TABLE IF EXISTS tempdb.title_basics_ratings;
CREATE TEMPORARY TABLE IF NOT EXISTS tempdb.title_basics_ratings
SELECT A.*, B.averageRating, B.numVotes, B.trid_1
FROM tempdb.title_basics_1 A
LEFT JOIN tempdb.title_ratings_1 B
ON A.tbid_1 = B.trid_1;

ALTER TABLE tempdb.name_basics ADD INDEX `idx011` (title_01);
ALTER TABLE tempdb.name_basics ADD INDEX `idx012` (title_02);
ALTER TABLE tempdb.name_basics ADD INDEX `idx013` (title_03);
ALTER TABLE tempdb.name_basics ADD INDEX `idx014` (title_04);
ALTER TABLE tempdb.name_basics ADD INDEX `idx015` (title_05);
ALTER TABLE tempdb.name_basics ADD INDEX `idx016` (title_06);

DROP TEMPORARY TABLE IF EXISTS tempdb.name_title_1;
CREATE TEMPORARY TABLE IF NOT EXISTS tempdb.name_title_1
SELECT A.*, B.primaryTitle, B.OriginalTitle, B.AverageRating, B.numVotes, B.tconst AS matchedmovie
FROM tempdb.name_basics A
INNER JOIN tempdb.title_basics_ratings B
ON A.title_01 = B.trid_1;
```

Evaluating our join criteria:
Disposition Rate: 14.26% (Percentage of records in title basics which were joined to a record in name basics)
Match Rate: 100% (Percentage of records in name basics which were joined to a record in title basics)

DATA PLATFORM

Overview of tables derived from our joins: Title_Basics_Ratings

```
8 ● SELECT * FROM `tempdb`.`title_basics_ratings`;
```

100% 1:1

Result Grid Filter Rows: Search Export: Fetch rows:

tconst	titleType	primaryTitle	originalTitle	isAdult	startYear	endYear	runtimeMinut...	Genre1	Genre2	Genre3	tb_id	tbid_1	averageRating	numVotes
tt0000001	short	Carmencita	Carmencita	0	1894	\N	1	Documentary	Short	NULL	t0000001	90000001	5.7	1834
tt0000002	short	Le clown et ses chiens	Le clown et ses chiens	0	1892	\N	5	Animation	Short	NULL	t0000002	90000002	6	236
tt0000003	short	Pauvre Pierrot	Pauvre Pierrot	0	1892	\N	4	Animation	Comedy	Romance	t0000003	90000003	6.5	1594
tt0000004	short	Un bon bock	Un bon bock	0	1892	\N	12	Animation	Short	NULL	t0000004	90000004	6	153
tt0000005	short	Blacksmith Scene	Blacksmith Scene	0	1893	\N	1	Comedy	Short	NULL	t0000005	90000005	6.2	2410
tt0000006	short	Chinese Opium Den	Chinese Opium Den	0	1894	\N	1	Short	NULL	NULL	t0000006	90000006	5.2	158
tt0000007	short	Corbett and Courtney Before the Kinetograph	Corbett and Courtney Before the Kinetograph	0	1894	\N	1	Short	Sport	NULL	t0000007	90000007	5.4	755
tt0000008	short	Edison Kinetoscopic Record of a Sneeze	Edison Kinetoscopic Record of a Sneeze	0	1894	\N	1	Documentary	Short	NULL	t0000008	90000008	5.5	1982
tt0000009	short	Miss Jerry	Miss Jerry	0	1894	\N	40	Romance	Short	NULL	t0000009	90000009	5.9	191
tt0000010	short	Leaving the Factory	La sortie de l'usine Lumière à Lyon	0	1895	\N	1	Documentary	Short	NULL	t0000010	90000010	6.9	6603
tt0000011	short	Akrobatisches Potpourri	Akrobatisches Potpourri	0	1895	\N	1	Documentary	Short	NULL	t0000011	90000011	5.2	331
tt0000012	short	The Arrival of a Train	L'arrivée d'un train à La Ciotat	0	1896	\N	1	Documentary	Short	NULL	t0000012	90000012	7.5	11325
tt0000013	short	The Photographical Congress Arrives in Lyon	Le débarquement du congrès de photograp...	0	1895	\N	1	Documentary	Short	NULL	t0000013	90000013	5.8	1747
tt0000014	short	The Waterer Watered	L'arroseur arrosé	0	1895	\N	1	Comedy	Short	NULL	t0000014	90000014	7.1	5086
tt0000015	short	Autour d'une cabine	Autour d'une cabine	0	1894	\N	2	Animation	Short	NULL	t0000015	90000015	6.2	960
tt0000016	short	Boat Leaving the Port	Barque sortant du port	0	1895	\N	1	Documentary	Short	NULL	t0000016	90000016	5.9	1347
tt0000017	short	Italienischer Bauerntanz	Italienischer Bauerntanz	0	1895	\N	1	Documentary	Short	NULL	t0000017	90000017	4.6	295
tt0000018	short	Das boxende Känguruh	Das boxende Känguruh	0	1895	\N	1	Short	NULL	NULL	t0000018	90000018	5.3	544
tt0000019	short	The Clown Barber	The Clown Barber	0	1898	\N	\N	Comedy	Short	NULL	t0000019	90000019	5.3	29
tt0000020	short	The Derby 1895	The Derby 1895	0	1895	\N	1	Documentary	Short	Sport	t0000020	90000020	5	319
tt0000022	short	Blacksmith Scene	Les forgerons	0	1895	\N	1	Documentary	Short	NULL	t0000022	90000022	5.1	1002
tt0000023	short	The Sea	Baignade en mer	0	1895	\N	1	Documentary	Short	NULL	t0000023	90000023	5.7	1290
tt0000024	short	Opening of the Kiel Canal	Opening of the Kiel Canal	0	1895	\N	\N	News	Short	NULL	t0000024	90000024	4.4	83
tt0000025	short	The Oxford and Cambridge University Boat...	The Oxford and Cambridge University Boat...	0	1895	\N	\N	News	Short	Sport	t0000025	90000025	4.3	39
tt0000026	short	The Messers. Lumière at Cards	Partie d'écarté	0	1896	\N	1	Documentary	Short	NULL	t0000026	90000026	5.7	1433
tt0000027	short	Cordeliers' Square in Lyon	Place des Cordeliers à Lyon	0	1895	\N	1	Documentary	Short	NULL	t0000027	90000027	5.6	1051
tt0000028	short	Fishing for Goldfish	La pêche aux poissons rouges	0	1895	\N	1	Documentary	Short	NULL	t0000028	90000028	5.2	973

Result Grid
Form Editor
Field Types
Query Stats
Execution Plan

DATA PLATFORM

Creating a join between different files:
IMDB API Data
Data on Movies different streaming platforms

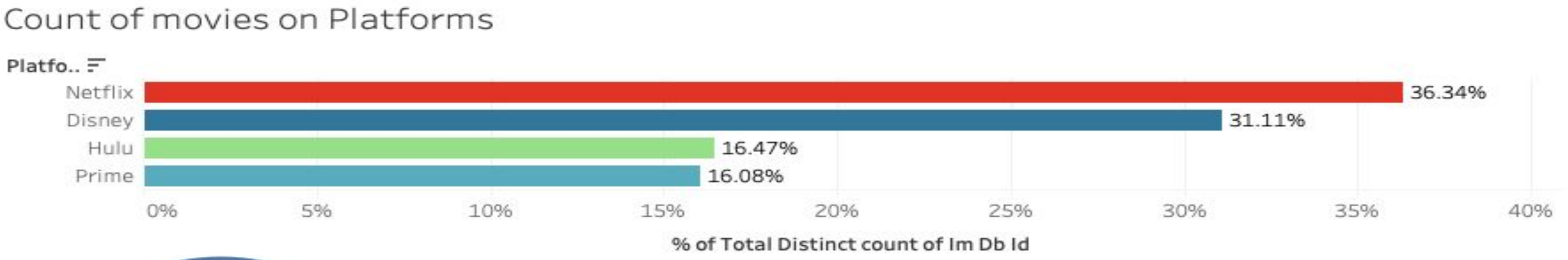
```sql
SELECT * FROM imdb.imdb_api_data;
SELECT * FROM imdb.imdb_api_data AS a
INNER JOIN imdb.moviesonstreamingplatforms_updated1 AS b
ON a.title = b.Title;
```

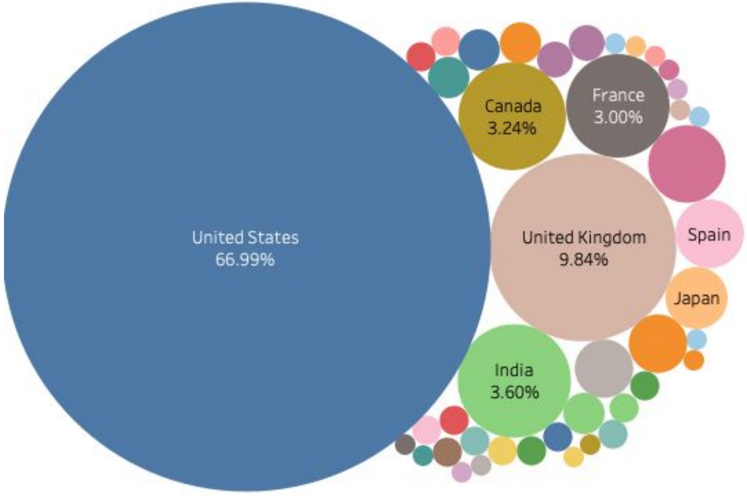
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

imDbId	title	fullTitle	type	year	imDb	metacritic	theMovieDb	rottenTomatoes	tV_com	filmAffinity	errorMessage	ID	Ti
tt1302006	The Irishman	The Irishman (2019)	Movie	2019	7.8	94	7.7	95		7.3		1	Th
tt5074352	Dangal	Dangal (2016)	Movie	2016	8.4		8			7.4		2	Da
tt11989890	David Attenborough: A Life on ...	David Attenborough: A Life on ...	Movie	2020	9	72	8.6	95		8.1		3	Da
tt0169102	Lagaan: Once Upon a Time in I...	Lagaan: Once Upon a Time in I...	Movie	2001	8.1	84	7.4	95		7		4	Lag
tt11464826	The Social Dilemma	The Social Dilemma (2020)	Movie	2020	7.6	78	7.5	85		6.8		7	Th
tt3967856	Okja	Okja (2017)	Movie	2017	7.3	75	7.5			6.6		8	Ok
tt6412452	The Ballad of Buster Scruggs	The Ballad of Buster Scruggs (2...	Movie	2018	7.3	79	7.2	89		6.5		9	Th
tt1070874	The Trial of the Chicago 7	The Trial of the Chicago 7 (2020)	Movie	2020	7.8	76	7.8	89		7.1		10	Th
tt10324144	Article 15	Article 15 (2019)	Movie	2019	8.2		7.7	90		6.9		11	Art
tt8526872	Dolemite Is My Name	Dolemite Is My Name (2019)	Movie	2019	7.3	76	7.1	97		6.3		13	Do
tt2396589	Mudbound	Mudbound (2017)	Movie	2017	7.4	85	7.5	97		6.7		14	Mu
tt0367110	Swades	Swades (2004)	Movie	2004	8.2		7.4	83		6.7		15	Sw
tt9412098	Fyre	Fyre (2019)	Movie	2019	7.2	75	6.9	92		6.5		16	Fyr
tt11388580	Miss Americana	Miss Americana (2020)	Movie	2020	7.4	65	7.9	91		6.2		17	Mis
tt3455224	Virunga	Virunga (2014)	Movie	2014	8.2	95	8			7.8		18	Vir
tt4934950	Talvar	Talvar (2015)	Movie	2015	8.2		7.6			6.4		20	Tal
tt7984766	The King	The King (2019)	Movie	2019	7.2	62	7.2	71		6.4		21	Th

ANALYTICS & INSIGHTS

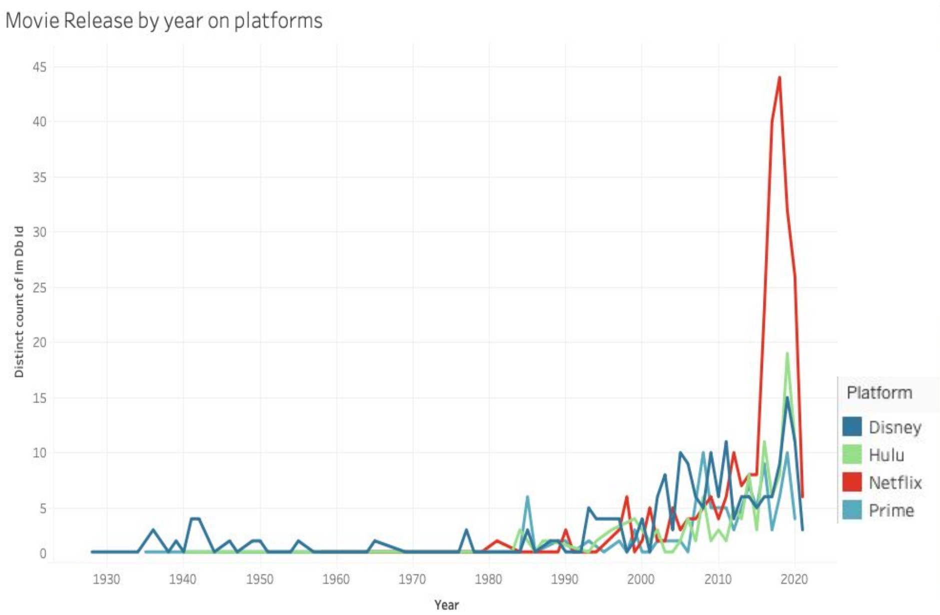
EDA

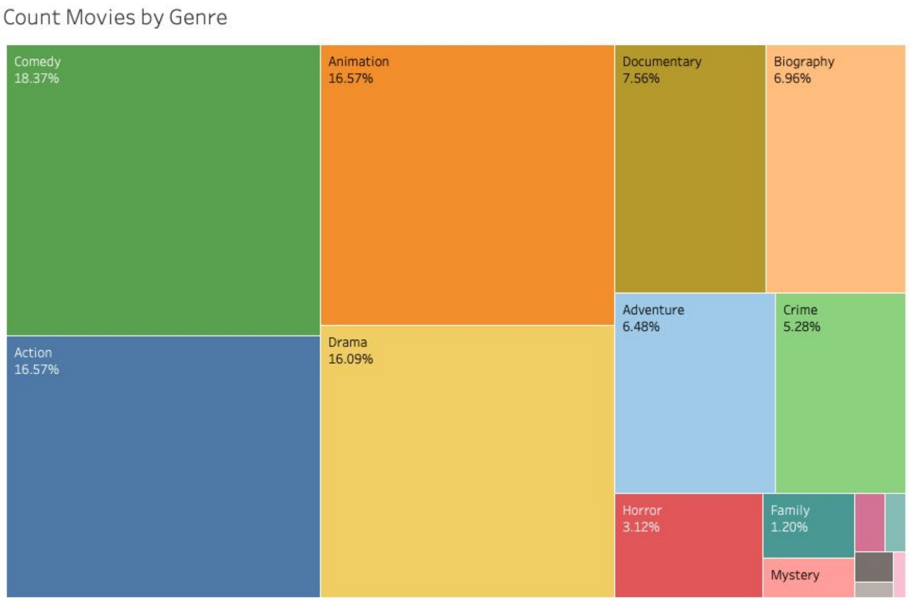




- Netflix hosts the highest share of movies.
- USA produces the highest share of movies, followed by UK and then India.

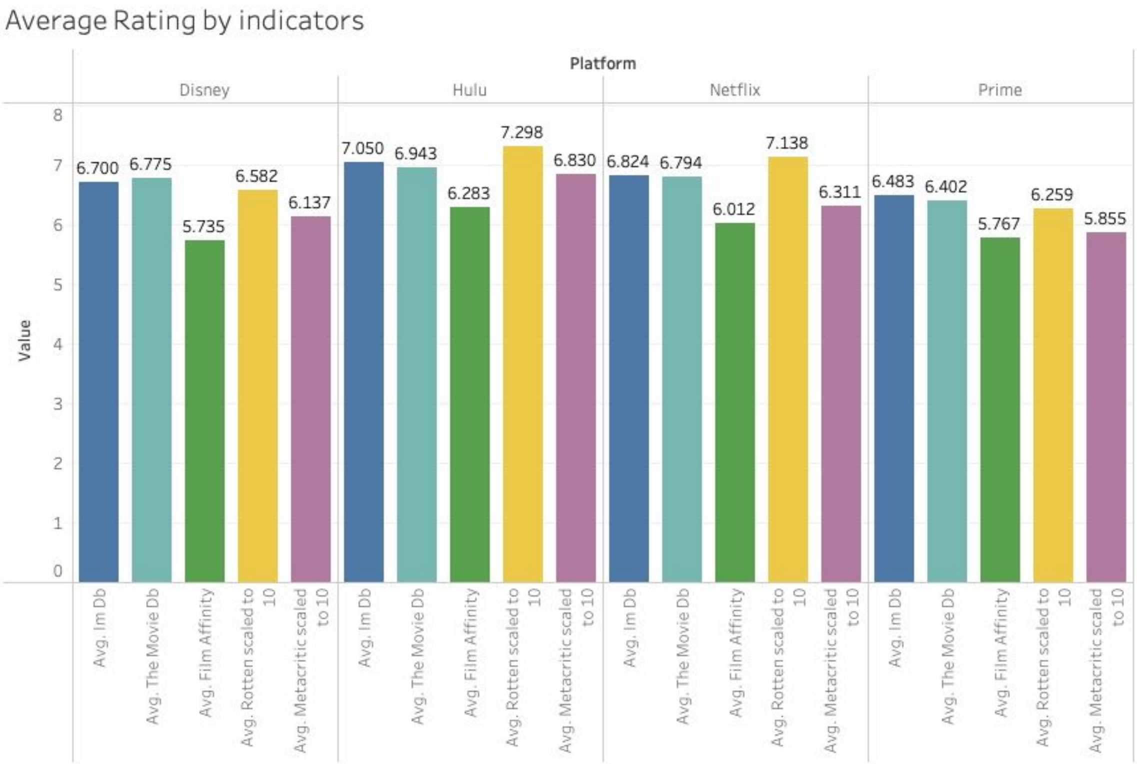
EDA





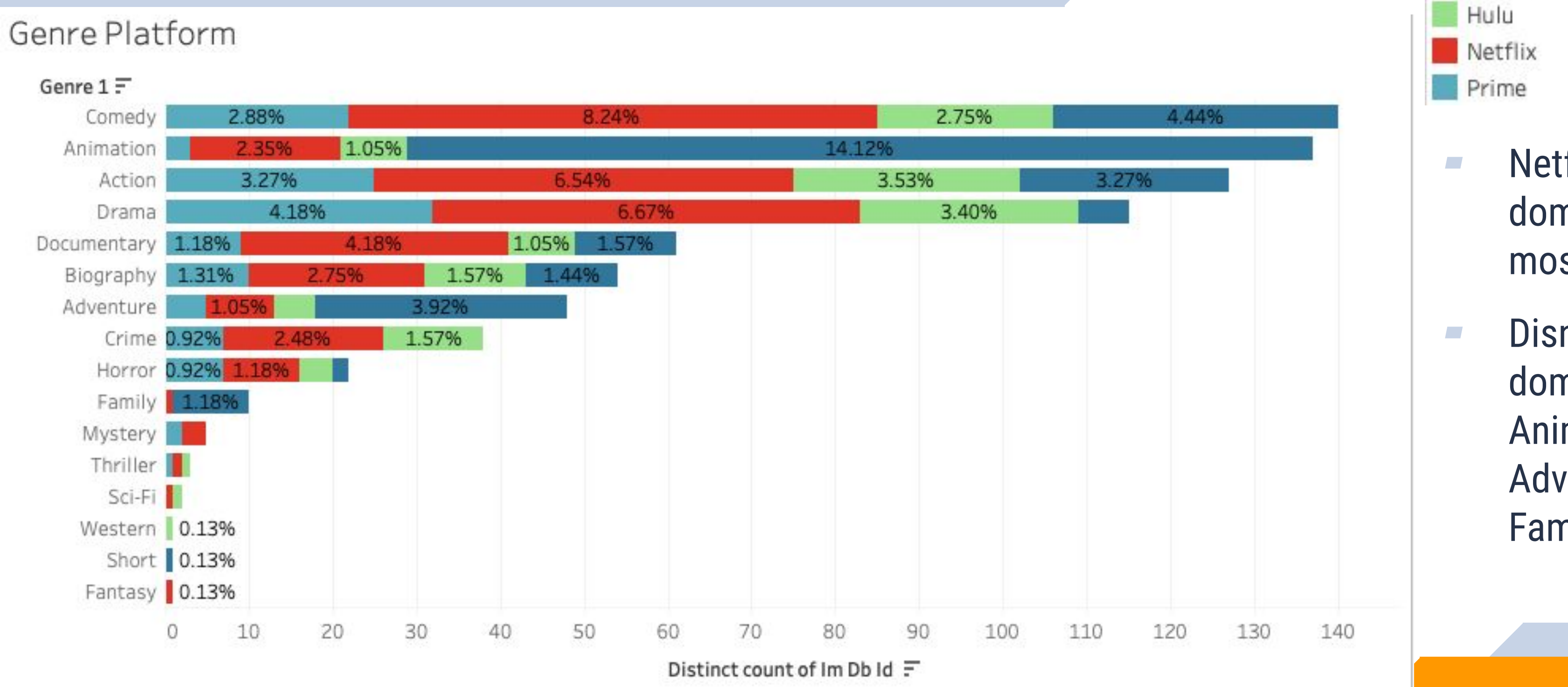
- Maximum movies released in 2019, acquired by Netflix & Hulu
- Comedy, Action, Animation and Drama top genres of movies produced

RATINGS INSIGHTS



- Rotten Tomatoes & Metacritic ratings had to be scaled to 10
- Movies on Hulu and Netflix rated higher across indicators, while movies on Prime and Disney are rated lower.
- Rating of Rotten Tomatoes is higher overall, while FilmAffinity rates lower

PLATFORM INSIGHTS



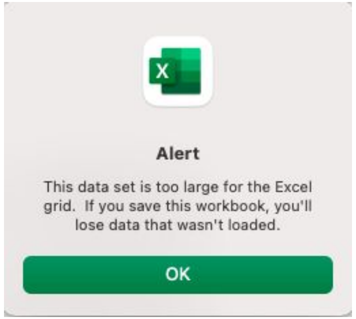
- Netflix dominates in most genres
- Disney dominates in Animation, Adventure and Family

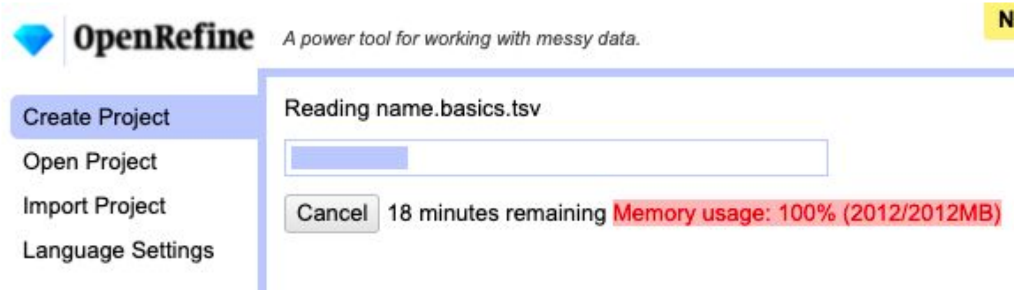
Interactive Dashboard

- Play video for an example of the interactive dashboard.
- Allows user to customise their filters and find hidden gems

PROBLEMS FACED, SOLUTIONS IMPLEMENTED & LESSONS LEARNED

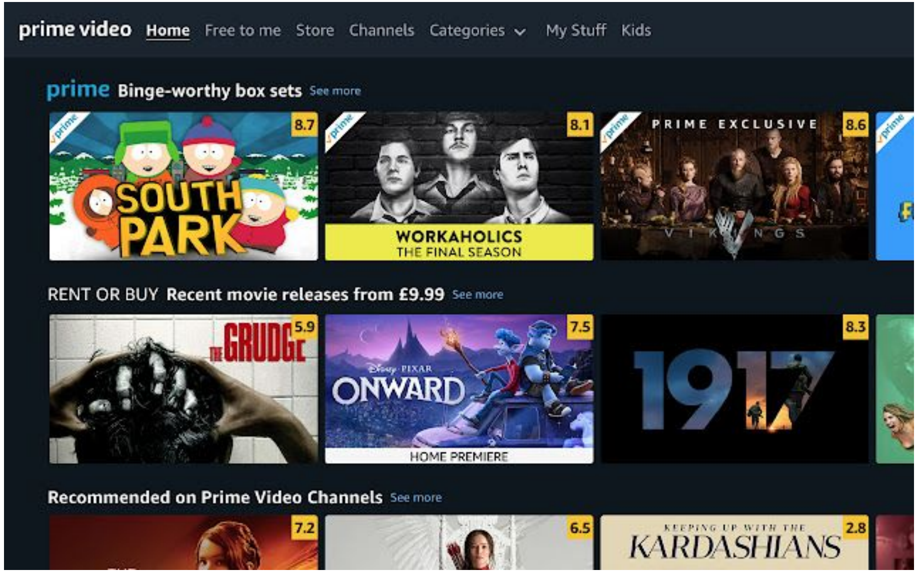
- **Data Collection:** Due to limits on API calls per user per day, we need to budget for more time and spread the web-scraping process across multiple users. However, we subscribed to a paid pricing plan.
- Local computers have memory limits, which can be solved on Cloud platform
- Excel cannot handle large data sets, while Python works
- **ETL Process:** We were unable to load data directly into MySQL due to an ASCII error. We used a 30 day trial period for Datagrip to import data.
- **Data Analysis:** Data type of measurement needs to be modified in Tableau based on use case





What Next?

Develop an application for users based on the interactive filter seen earlier



Prime video currently shows ImDB ratings for its movies

Pitch the idea to link movie ratings to other platforms

THANK YOU

QUESTIONS?