



# **STREAMING SERVICE ANALYSIS - AZURE SQL**

**ABHISHEK TIKAM RAMCHANDANI - 002743745**

**JURREYAH FIRDAWS MOHAMMED - 002747514**

**SAMHITHA MEREDDY - 002796140**

**ATHARVA UPLENCHWAR - 002990536**

# CONTENT

**01**

ARCHITECTURE DIAGRAM

**02**

RELATIONAL DIAGRAM

**03**

DATA CLEANING

**04**

PIPELINING THE DATA

**05**

DATA REFRESH

**06**

VISUALIZATIONS

**07**

Q&A

# Objectives

01

02

03

04

05

06

The project plans to focus on creating a relational database for storing and retrieving user, show, episode, subscription, and view information for streaming services.

Analyze user behavior to discover trends, viewing preferences, and watching patterns.

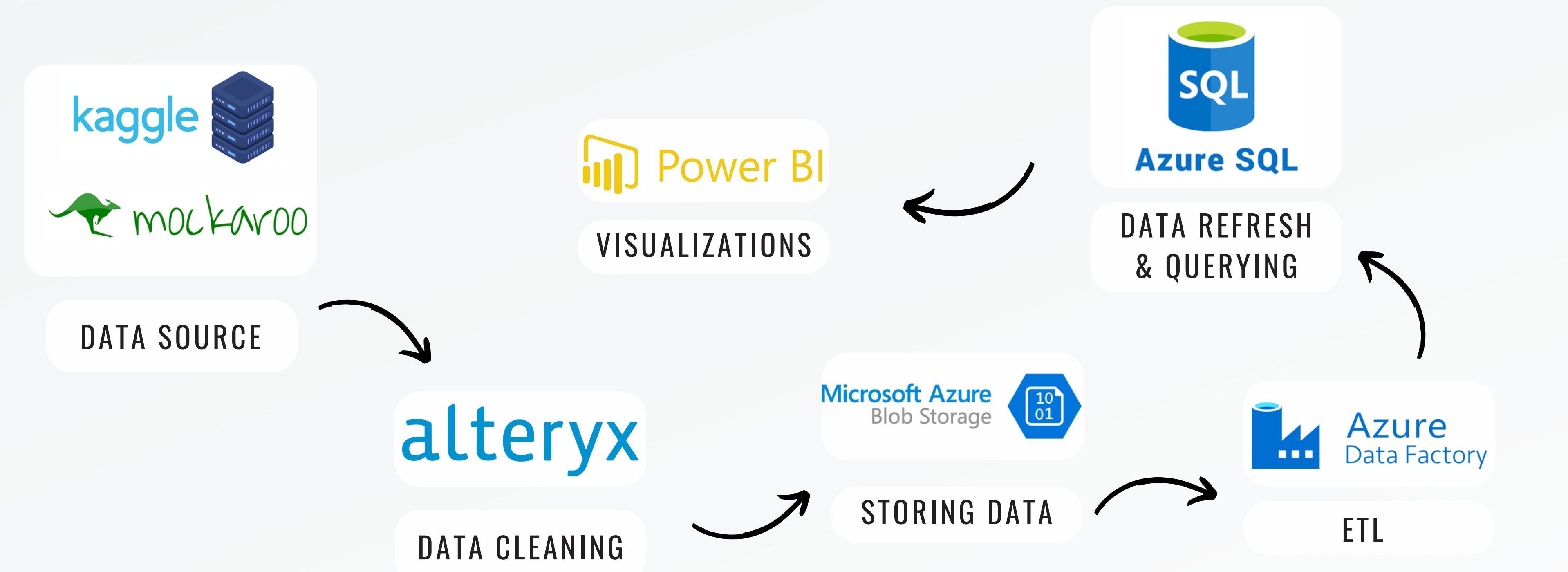
Determine the level of popularity for various forms of media, such as films, TV series, genres, and actors.

Offer streaming service providers business insights for selecting content licenses, devising ad campaigns, and enhancing user engagement.

In order to generate useful insights, the project will concentrate on analyzing data from streaming providers such as Netflix, Hulu, Disney+ etc.

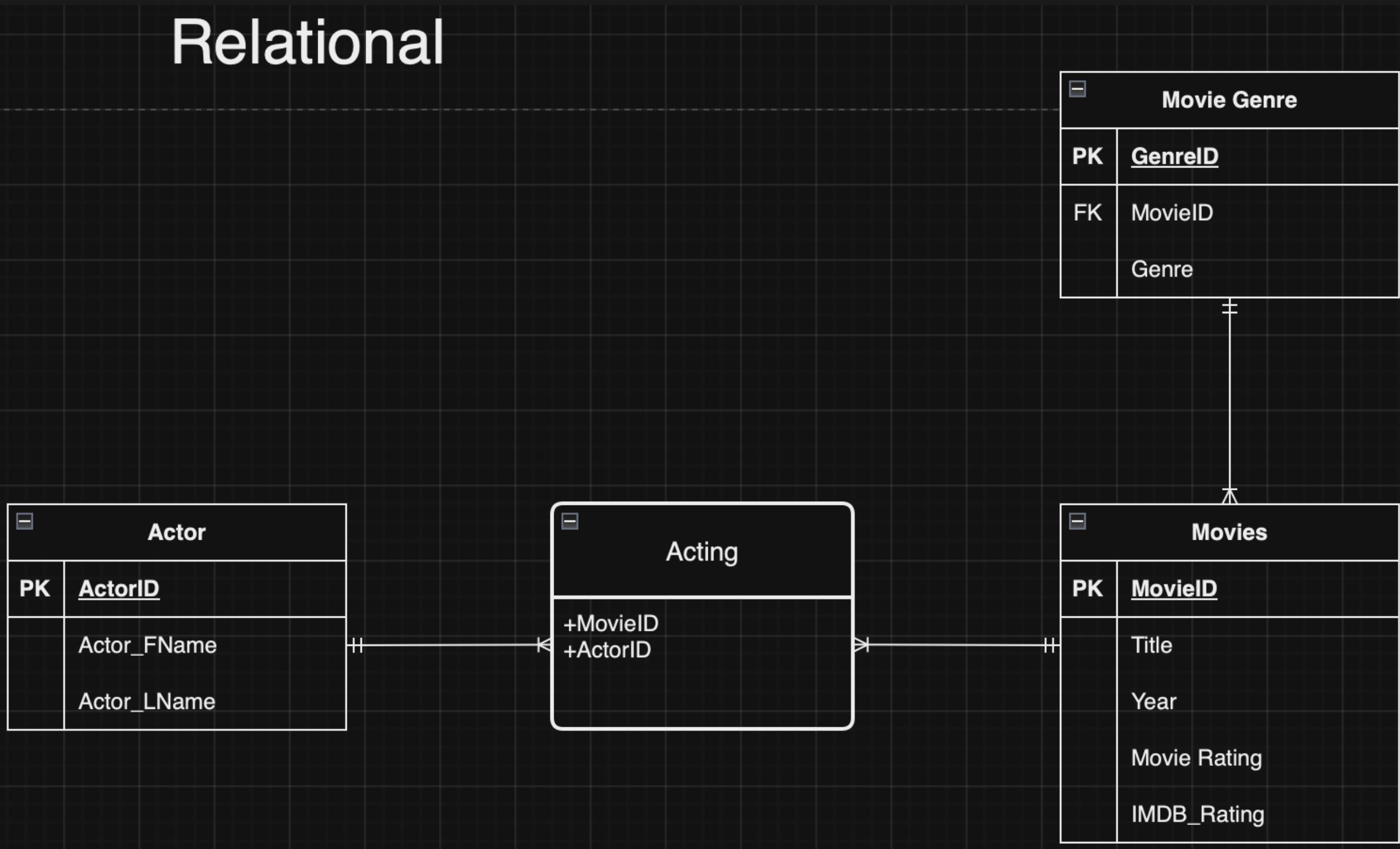
Reports and data visualizations will be produced to highlight key findings for stakeholders.

# ARCHITECTURE DIAGRAM



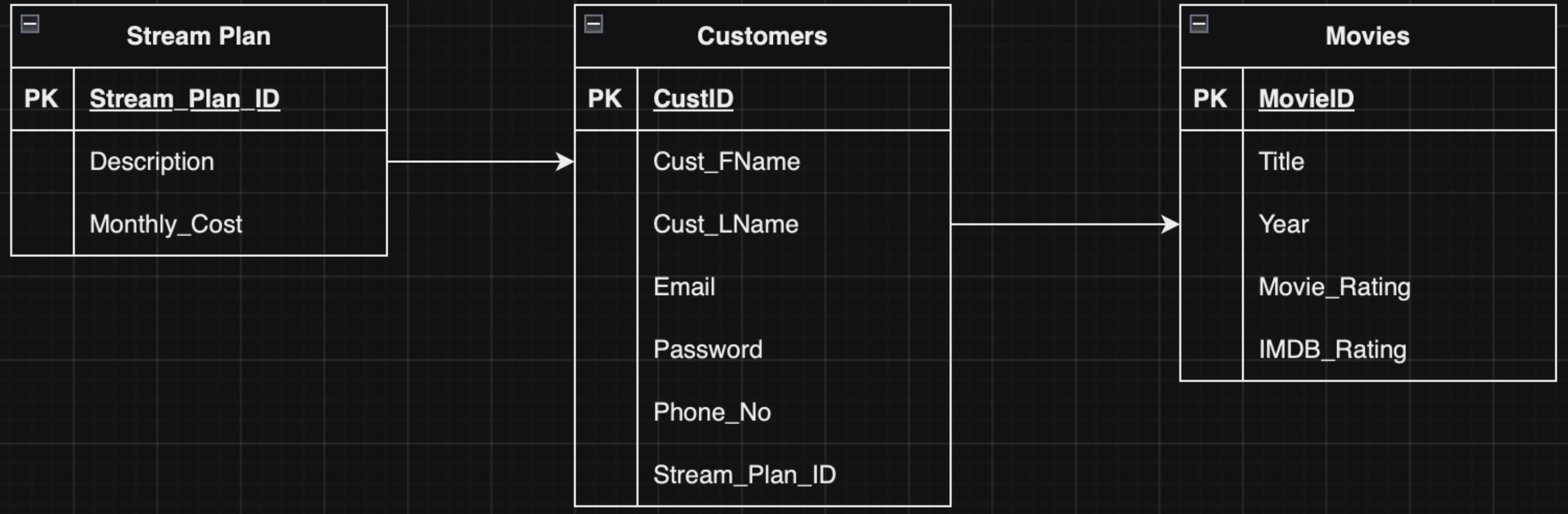
# ERD-RELATIONAL DATABASE

Relational



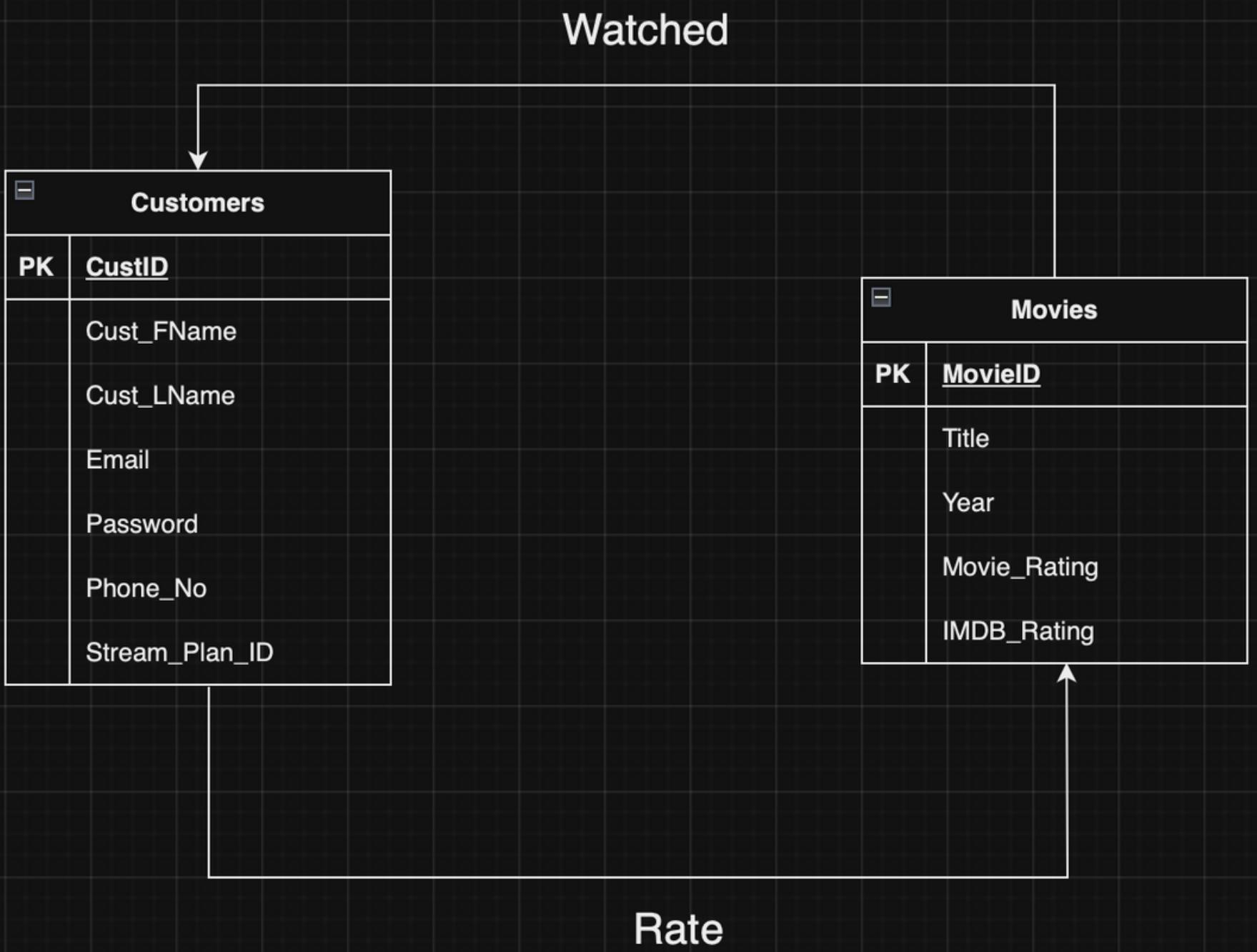
# ERD - DOCUMENT DATABASE

## Document



# ERD-GRAPH DATABASE

## Graph



# RAW-DATASET

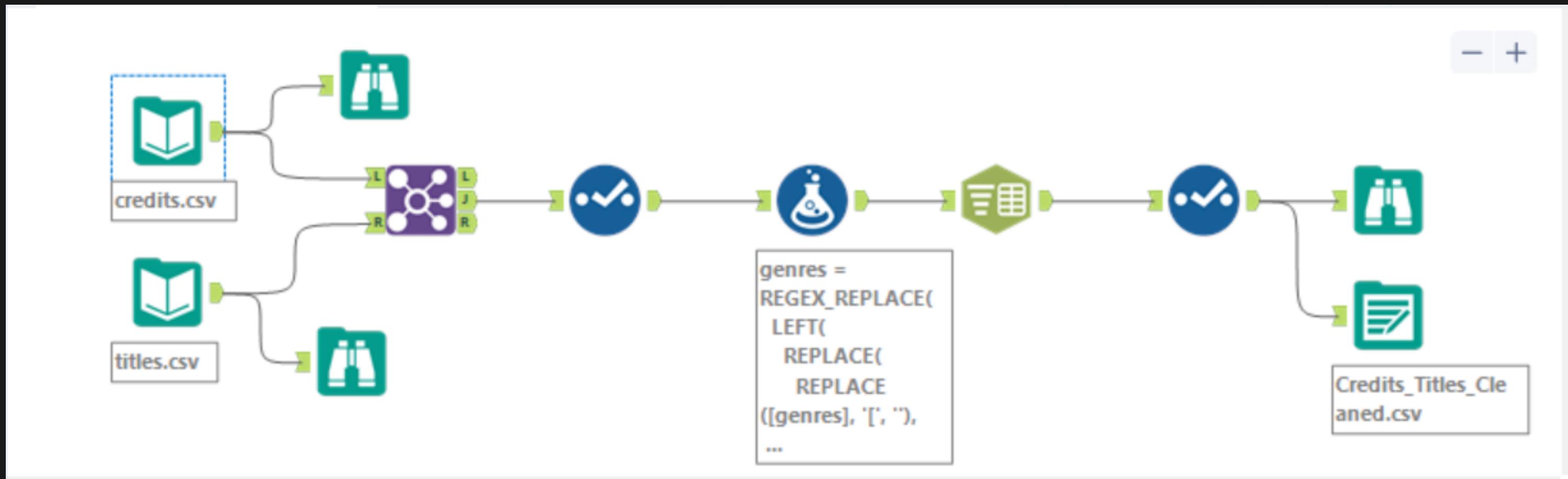
id	title	type	description	release_year	age_certif	runtime	genres	production_companies	seasons	imdb_id	imdb_score	imdb_votes	imdb_popularity	tmdb_score
ts300399	Five Came Back: The Referendum	SHOW	This collection includes five films from the 1940s and 1950s that have been restored and re-scored by Academy Award-winning composer Hans Zimmer.	1945	TV-MA	51	['documentary', 'war']	['US']	1				0.6	
tm84618	Taxi Driver	MOVIE	A mentally unstable New York City taxi driver observes his world through a distorted lens.	1976	R	114	['drama', 'thriller']	['US']		tt0075314	8.2	808582	40.965	8.179
tm154986	Deliverance	MOVIE	Intent on seeing justice done, two men travel to the Georgia mountains to rescue a woman held captive by her sadistic husband.	1972	R	109	['drama', 'thriller']	['US']		tt0068473	7.7	107673	10.01	7.3
tm127384	Monty Python and the Holy Grail	MOVIE	King Arthur, accompanied by his Knights of the Round Table, sets off on a quest to find the Holy Grail.	1975	PG	91	['fantasy', 'comedy']	['GB']		tt0071853	8.2	534486	15.461	7.811
tm120801	The Dirty Dozen	MOVIE	12 American military veterans are recruited to carry out a dangerous mission behind enemy lines.	1967		150	['war', 'action']	['GB', 'US']		tt0061578	7.7	72662	20.398	7.6
ts22164	Monty Python's Flying Circus	SHOW	A British sketch comedy series featuring the members of the comedy group Monty Python.	1969	TV-14	30	['comedy']	['GB']	4	tt0063929	8.8	73424	17.617	8.306
tm70993	Life of Brian	MOVIE	Brian Cohen is a man who becomes involved in the life of Jesus Christ.	1979	R	94	['comedy']	['GB']		tt0079470	8	395024	17.77	7.8
tm14873	Dirty Harry	MOVIE	When a madman begins killing police officers, San Francisco's most famous cop goes after him.	1971	R	102	['thriller']	['US']		tt0066999	7.7	155051	12.817	7.5
tm119281	Bonnie and Clyde	MOVIE	In the 1930s, bank robbers Bonnie Parker and Clyde Barrow become fugitives.	1967	R	110	['crime']	['US']		tt0061418	7.7	112048	15.687	7.5
tm98978	The Blue Lagoon	MOVIE	Two small children are stranded on a remote island.	1980	R	104	['romance']	['US']		tt0080453	5.8	69844	50.324	6.156
tm44204	The Guns of Navarone	MOVIE	A team of allied soldiers must destroy four naval guns on a Greek island.	1961		158	['action']	['GB', 'US']		tt0054953	7.5	50748	13.844	7.3
tm67378	The Professionals	MOVIE	An arrogant Texan and his partner plan to rob a bank.	1966	PG-13	117	['western']	['US']		tt0060862	7.3	16446	13.123	7.1
tm69997	Richard Pryor: Live in Concert	MOVIE	Richard Pryor delivers a stand-up comedy performance.	1979	R	78	['comedy']	['US']		tt0079807	8.1	5141	4.718	7.5
tm16479	White Christmas	MOVIE	Two talented songwriters are recruited to perform at a hotel during the holidays.	1954		115	['romance']	['US']		tt0047673	7.5	42488	8.915	7.2
tm135083	Cairo Station	MOVIE	Qinawi, a physician, is recruited to work in Cairo.	1958		77	['drama']	['EG']		tt0051390	7.5	4471	5.546	7.3
tm89386	Hitler: A Career	MOVIE	A keen chronicler of Hitler's rise to power.	1977	PG	150	['history']	['DE']		tt0191182	7.5	2460	4.305	7.3
tm156453	FTA	MOVIE	A documentary about the Free Trade Area of the Americas.	1972	R	97	['war', 'documentary']	['US']		tt0068562	6.2	418	1.268	6.1

**Titles Table**

person_id	id	name	character	role
3748	tm84618	Robert De Niro	Travis Bickle	ACTOR
14658	tm84618	Jodie Foster	Iris Steensma	ACTOR
7064	tm84618	Albert Brooks	Tom	ACTOR
3739	tm84618	Harvey Keitel	Matthew 'Sport' Higgins	ACTOR
48933	tm84618	Cybill Shepherd	Betsy	ACTOR
32267	tm84618	Peter Boyle	Wizard	ACTOR
519612	tm84618	Leonard Harris	Senator Charles Palantine	ACTOR
29068	tm84618	Diahnne Abbott	Concession Girl	ACTOR
519613	tm84618	Gino Arditto	Policeman at Rally	ACTOR
3308	tm84618	Martin Scorsese	Passenger Watching Silhouette	ACTOR
43791	tm84618	Murray Moston	Iris' Time Keeper	ACTOR
519614	tm84618	Richard Higgs	Secret Service Agent	ACTOR
519615	tm84618	Bill Minkin	Tom's Assistant (uncredited)	ACTOR

**Credits Table**

# ALTERYX - CLEANING



age_certification	PG-13	
fx	IF ISNULL([age_certification]) THEN 'PG-13' ELSE [age_certification] ENDIF	
X		
Data type:	V_String	
Size:	254	

Select Column to Split

Column to split: name

Delimiters: (empty)

Split to columns

Number of columns: 2

Extra characters: Leave extra in last column

Output root name: name

	Field	Type	Size	Rename
	✓ id	V_String	254	
	✓ name	V_String	254	
	✓ title	V_String	254	
	✓ release_year	V_String	254	
	✓ age_certification	V_String	254	
	✓ genres	V_String	254	
	✓ imdb_score	Float	4	
	✓ name1	V_String	254	first_name
	✓ name2	V_String	254	last_name
	✓ *Unknown	Unknown	0	

# DATASET AFTER CLEANING

id	title	type	description	release_year	age_certif	runtime	genres	production_companies	seasons	imdb_id	imdb_score	imdb_votes	tmdb_popularity	tmdb_score
ts300399	Five Came Back: The Referer	SHOW	This collection i	1945	TV-MA	51	['documentary', 'historical']		1				0.6	
tm84618	Taxi Driver	MOVIE	A mentally unst	1976	R	114	['drama', 'psychological']			tt0075314	8.2	808582	40.965	8.179
tm154986	Deliverance	MOVIE	Intent on seeing	1972	R	109	['drama', 'adventure']			tt0068473	7.7	107673	10.01	7.3
tm127384	Monty Python and the Holy	MOVIE	King Arthur, acc	1975	PG	91	['fantasy', 'comedy']			tt0071853	8.2	534486	15.461	7.811
tm120801	The Dirty Dozen	MOVIE	12 American mi	1967		150	['war', 'action']			tt0061578	7.7	72662	20.398	7.6
ts22164	Monty Python's Flying Circus	SHOW	A British sketch	1969	TV-14	30	['comedy']		4	tt0063929	8.8	73424	17.617	8.306
tm70993	Life of Brian	MOVIE	Brian Cohen is a	1979	R	94	['comedy']			tt0079470	8	395024	17.77	7.8
tm14873	Dirty Harry	MOVIE	When a madman	1971	R	102	['thriller', 'action']			tt0066999	7.7	155051	12.817	7.5
tm119281	Bonnie and Clyde	MOVIE	In the 1930s, b	1967	R	110	['crime', 'drama']			tt0061418	7.7	112048	15.687	7.5
tm98978	The Blue Lagoon	MOVIE	Two small child	1980	R	104	['romance']			tt0080453	5.8	69844	50.324	6.156

**Titles**

**Uncleaned**

id	Title	Year	Movie_Rating	Genre	Ratings
ts300399	NocturnalAnimals	2016	R	thriller	7.5
tm84618	Dharmakshetra	2014	TV-14	drama	8.3
tm154986	BoBurnham:Inside	2021	R	comedy	8.7
tm127384	RiseofEmpires:Ottoman	2020	PG-13	drama	7.9
tm120801	IO	2019	PG-13	scifi	4.7
ts22164	AndBreatheNormally	2018	PG-13	drama	6.9
tm70993	PutYourHeadonMyShoulder	2019	TV-Y7	drama	8
tm14873	RealRob	2015	TV-14	comedy	6.3
tm119281	TheDUFF	2015	PG-13	comedy	6.4
tm98978	TheLincolnLawyer	2022	TV-MA	crime	7.7
tm44204	YooHoototheRescue	2019	TV-Y	animation	6.8
tm67378	DI4RI	2022	PG-13	family	7.3
tm69997	AjabPremKiGhazabKahani	2009	PG	comedy	6.3
tm16479	RaginiMMS	2011	PG-13	drama	5.1
tm135083	Kaaval	2021	PG-13	thriller	5.1
tm89386	WelcometoWeddingHell	2022	TV-14	drama	6.9
tm156453	IntothInferno	2016	PG-13	documentation	7.2

**Titles**  
**Cleaned**

# DATASET AFTER CLEANING

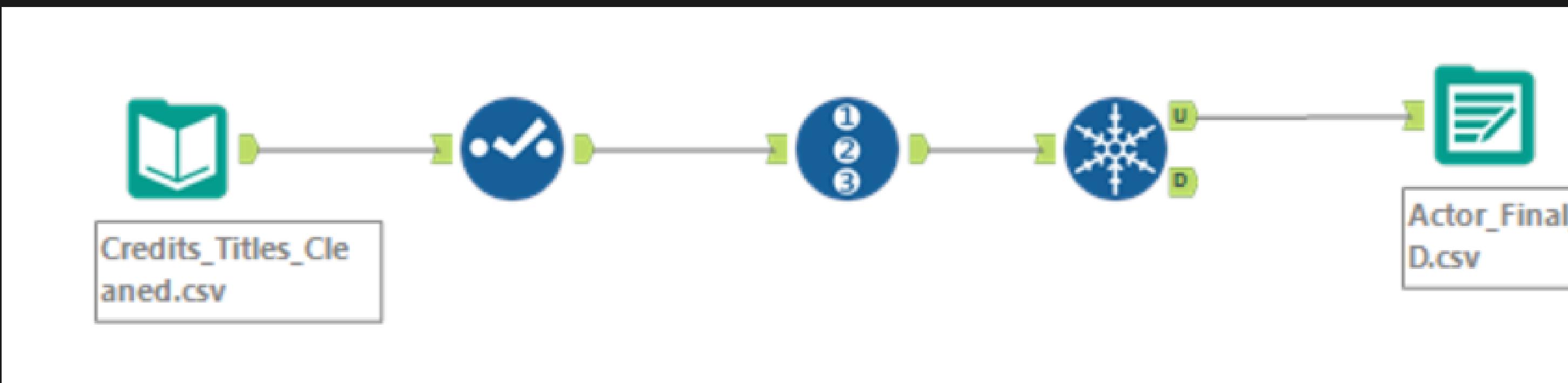
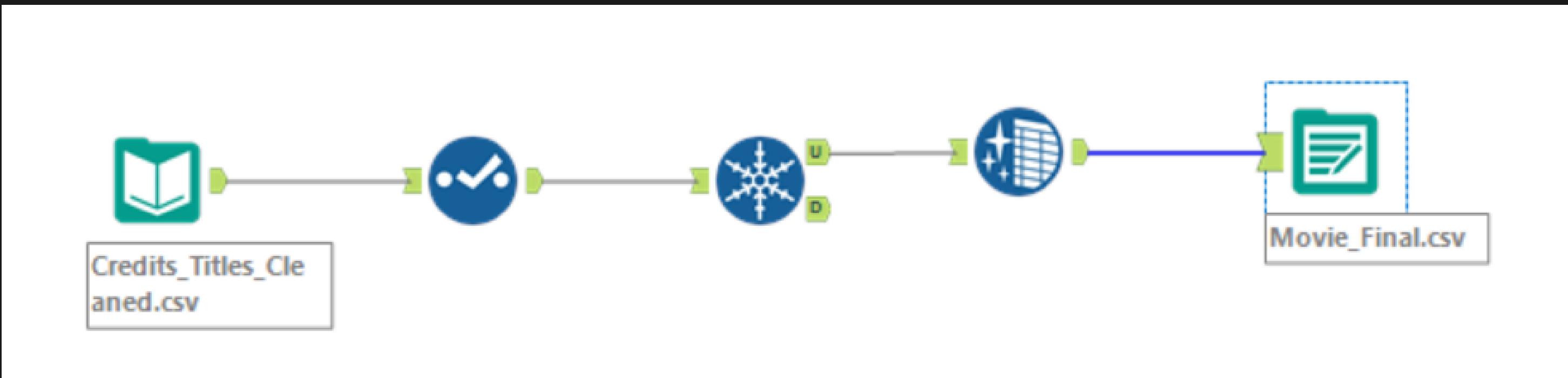
## Credits Uncleaned

person_id	id	name	character	role
3748	tm84618	Robert De Niro	Travis Bickle	ACTOR
14658	tm84618	Jodie Foster	Iris Steensma	ACTOR
7064	tm84618	Albert Brooks	Tom	ACTOR
3739	tm84618	Harvey Keitel	Matthew 'Sport' Higgins	ACTOR
48933	tm84618	Cybill Shepherd	Betsy	ACTOR
32267	tm84618	Peter Boyle	Wizard	ACTOR
519612	tm84618	Leonard Harris	Senator Charles Palantine	ACTOR
29068	tm84618	Diahnne Abbott	Concession Girl	ACTOR
519613	tm84618	Gino Ardito	Policeman at Rally	ACTOR
3308	tm84618	Martin Scorsese	Passenger Watching Silhouette	ACTOR
43791	tm84618	Murray Moston	Iris' Time Keeper	ACTOR
519614	tm84618	Richard Higgs	Secret Service Agent	ACTOR
519615	tm84618	Bill Minkin	Tom's Assistant (uncredited)	ACTOR
82426	tm84618	Bob Maroff	Mafioso (uncredited)	ACTOR
20935	tm84618	Victor Argo	Melio, Delicatessen Owner	ACTOR
7753	tm84618	Joe Spinell	Personell Officer	ACTOR
43279	tm84618	Robinson Frank Adu	Angry Black Man (uncredited)	ACTOR

## Credits Cleaned

id	Actor_Fname	Actor_LName
tm84618	Robert	De Niro
tm84618	Jodie	Foster
tm84618	Albert	Brooks
tm84618	Harvey	Keitel
tm84618	Cybill	Shepherd
tm84618	Peter	Boyle
tm84618	Leonard	Harris
tm84618	Diahnne	Abbott
tm84618	Gino	Ardito
tm84618	Martin	Scorsese
tm84618	Murray	Moston
tm84618	Richard	Higgs
tm84618	Bill	Minkin
tm84618	Bob	Maroff
tm84618	Victor	Argo
tm84618	Joe	Spinell
tm84618	Robinson	Frank Adu

# ALTERYX DATA FLOW



# MOCK DATASET

## Customers Table

mockaroo

SCHEMAS DATASETS MOCK APIs SCENARIOS PROJECTS FUNCTIONS ⚙️ ? 🚤 UPGRADE NOW

Field Name	Type	Options
Cust_ID	Row Number	blank: 0 % $\Sigma$ X
first_name	First Name	blank: 0 % $\Sigma$ X
last_name	Last Name	blank: 0 % $\Sigma$ X
email	Email Address	blank: 0 % $\Sigma$ X
password	Password	minimum length: 8 upper: 1 lower: 1 numbers: 1 symbols: 1 blank: 0 % $\Sigma$ X
BirthDate	Datetime	01/01/1950 <input type="button" value="Calendar"/> to 11/28/2023 <input type="button" value="Calendar"/> format: mm/dd/yyyy <input type="button" value="▼"/> blank: 0 % $\Sigma$ X
Phone_No	Phone	format: ##### <input type="button" value="▼"/> blank: 0 % $\Sigma$ X
Stream_Plan_ID	Custom List	1,2,3 <input type="button" value="Random"/> random <input type="button" value="▼"/> blank: 0 % $\Sigma$ X

+ ADD ANOTHER FIELD

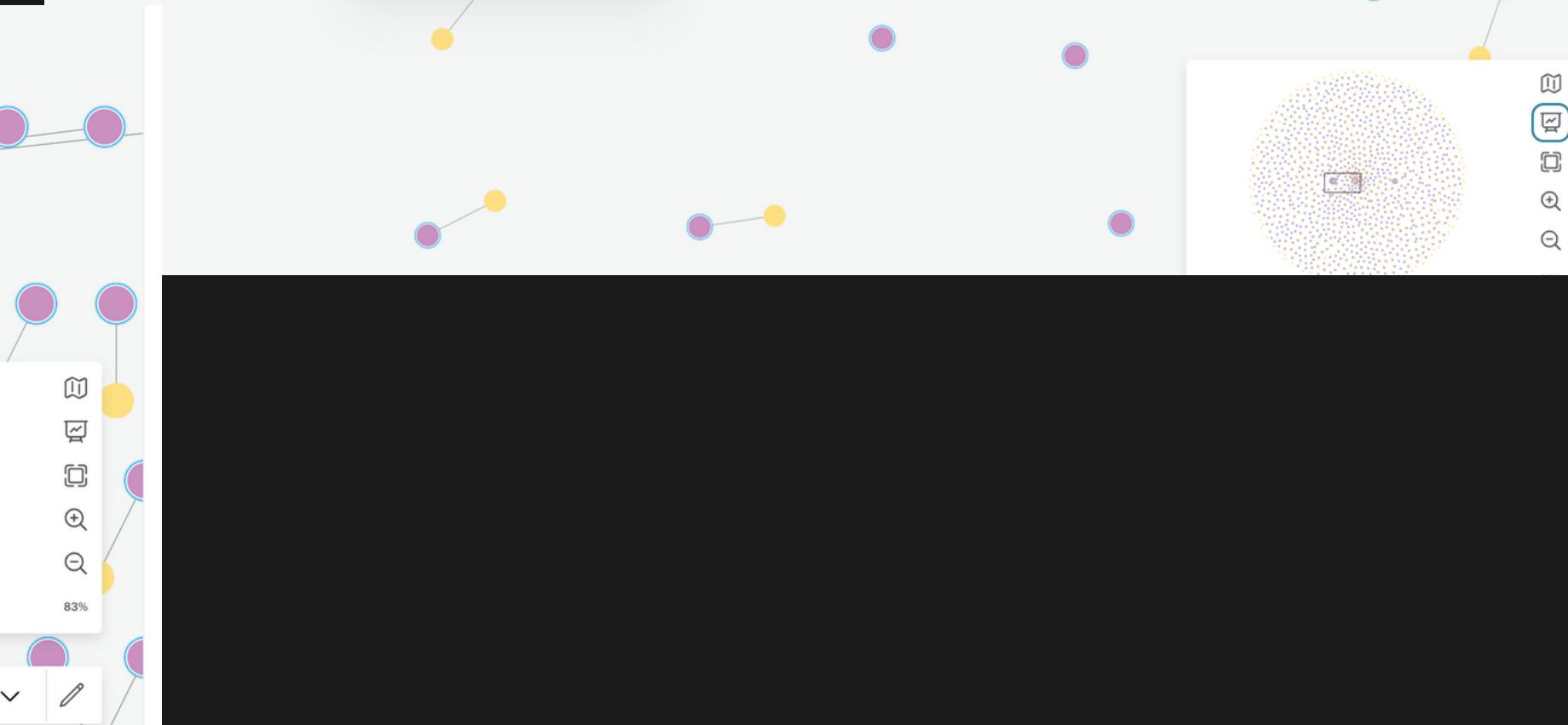
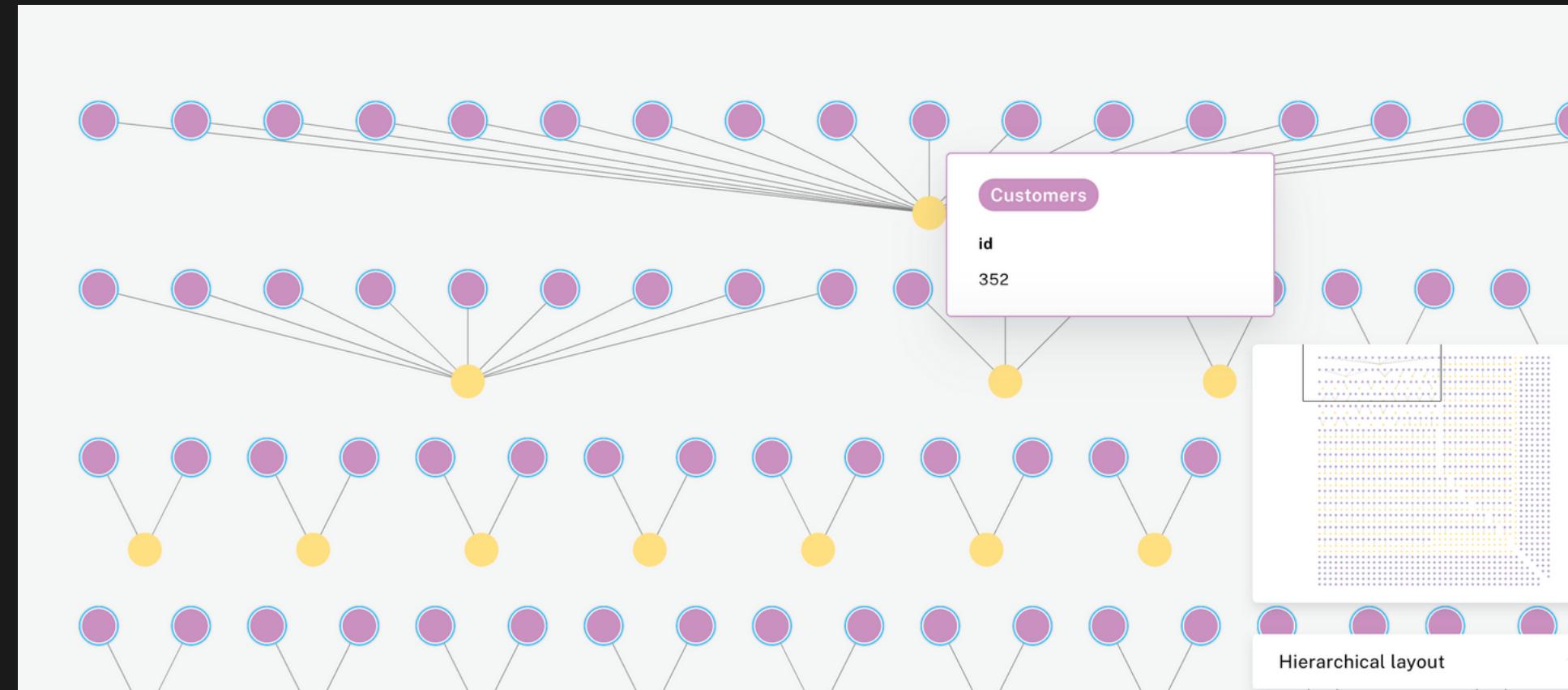
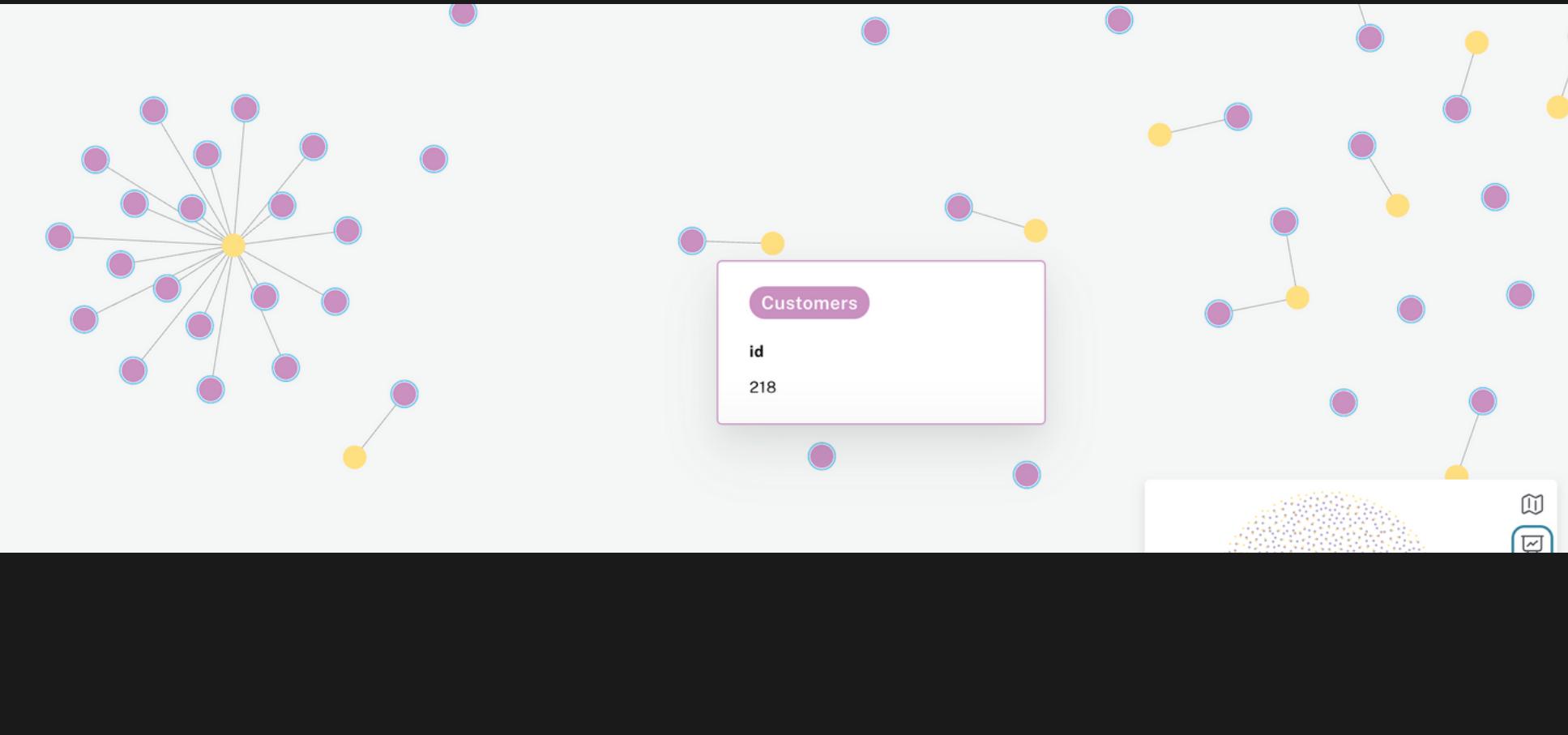
CustID	Cust_Fname	Cust_Lname	Email	Birthdate	Phone_No	Stream_Plan_ID
1	Goddard	Glassup	gglassup0@u	5/31/2016	9285113000	2
2	Vito	Newlands	vnewlands1@o	5/31/2019	5669242007	3
3	Bealle	Dobbinson	bdobbinson2@	8/13/1953	9021631876	1
4	Pat	Morton	pmorton3@o	9/27/1959	1832061480	3
5	Lucinda	Keneleysic	lkeneleyside4@	9/3/2018	9058120542	3

# Graph Model Implementation

Using Neo4J

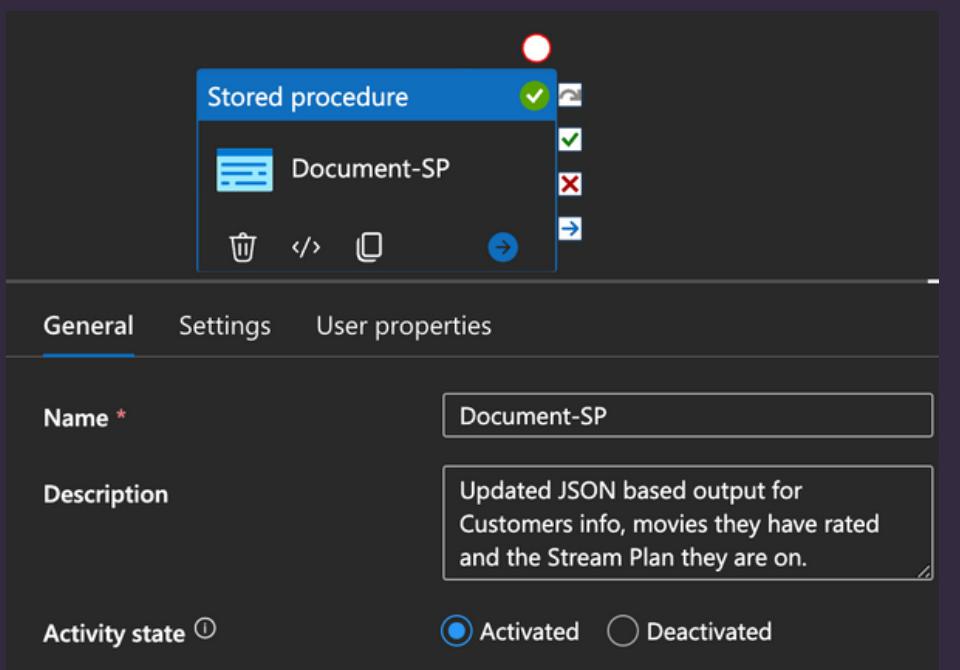
The screenshot illustrates the process of implementing a graph model using Neo4J. On the left, the Neo4J Neo4j Browser interface shows a configuration for importing data from a CSV file named "New\_Ratings.csv". The relationship type is set to "Rated", and the node ID mapping maps "Customers" to "id" and "Movies" to "Movie\_ID". A preview of the graph shows a single edge between a "Customers" node and a "Movies" node, labeled "Rated". On the right, the Neo4J Graph Data Grid visualization shows a network of nodes. Nodes are colored purple or yellow, representing different entities. A tooltip for a purple node displays the label "Customers" and the property "id" with the value "218". A zoomed-in view in the bottom right corner shows a cluster of nodes with a color-coded legend: purple, yellow, blue, and green. The bottom right corner also features a navigation sidebar with various icons.

# Graph Visualisations: Customers (Rated->) Movies



# DOCUMENT MODEL

```
CREATE PROCEDURE GetCustomerDetails
AS
BEGIN
    SELECT
        c.CustID AS id,
        c.Cust_FName AS firstName,
        c.Cust_LName AS lastName,
        c.Email AS email,
        sp.Description AS streamPlanDescription,
        sp.Monthly_Cost AS streamPlanMonthlyCost,
        (SELECT
            r.MovieID AS movieID,
            r.IMDB_Rating AS rating
        FROM Ratings r
        WHERE r.CustID = c.CustID
        FOR JSON PATH) AS ratedMovies
    FROM Customers c
    INNER JOIN Stream_Plan sp ON sp.Stream_Plan_ID = c.Stream_Plan_ID
    FOR JSON PATH;
END;
```



```
"id": 1,
"firstName": "Goddard",
"lastName": "Glassup",
"email": "gglassup0@un.org",
"streamPlanDescription": "Standard: Unlimited ad-free movies and TV shows, watch on 2 devices at a time, Full HD, download",
"streamPlanMonthlyCost": 15.49,
"ratedMovies": [
    {
        "movieID": "ts20358",
        "rating": 6.2
    },
    {
        "movieID": "ts273318",
        "rating": 5.9
    }
],
{
    "id": 2,
    "firstName": "Vito",
    "lastName": "Newlands",
    "email": "vnewlands1@e-recht24.de",
    "streamPlanDescription": "Premium: Unlimited ad-free movies and TV shows, watch on 4 devices at a time, Ultra HD, download",
    "streamPlanMonthlyCost": 22.99
},
{
    "id": 3,
    "firstName": "Bealle",
    "lastName": "Dobbinson",
    "email": "bdobbinson2@slashdot.org",
    "streamPlanDescription": "Standard with ads: Ad-supported, watch on 2 devices at a time, Full HD, download on 2 devices",
    "streamPlanMonthlyCost": 6.99
},
```

# DATA PIPELINING

▲ Pipelines 8

- blob BlobToSQLActing
- blob BlobToSQLActor
- blob BlobToSQLCustomers
- blob BlobToSQLGenre
- blob BlobToSQLMovies
- blob BlobToSQLRatings
- blob BlobToSQLStreamPlan
- blob Stored Procedure

Copy data ✓  
CustomersCopyData ✓

Parameters Variables Settings Output

Pipeline run ID: 2167d874-0aa5-4b21-ad16-ec9077f556f6 ⏪ ⓘ

Pipeline status ✓ Succeeded [View debug run consumption](#)

All status ▾ [Monitor in Azure Metrics](#) [Export to CSV](#) ▾

Showing 1 - 1 of 1 items

Activity name ↑↓	Activity status ↑↓	Activity type ↑↓	Run start ↑↓	Duration
CustomersCopyData ✓ Succeeded	Copy data	12/7/2023, 4:20:08 PM	15s	

# DATA REFRESH-RATINGS

The screenshot shows the Azure Data Factory studio interface. On the left, a sidebar lists various activities: StreamPlanSQLDB, BlobToSQLStreamP..., CustomersTable, CustomersSQLDB, and BlobToSQLCustom... . The 'Activities' section is expanded, showing options like 'Move and transform', 'Synapse', 'Azure Data Explorer', etc. In the center, a 'Copy data' activity is selected. The 'General' tab is active, showing the activity's name as 'RatingsCopyData'. The 'Source' tab shows the source as 'BlobContainer' and the 'Sink' tab shows the sink as 'SQLTable'. The 'Mapping' tab is visible. At the top, there are buttons for 'Validate', 'Validate copy runtime', 'Debug', and 'Trigger (1)'. A modal window titled 'Edit trigger' is open, showing the configuration for the 'RatingsTrigger'. The trigger is set to 'ScheduleTrigger' type, starting at '12/7/2023, 10:05:00 PM' in 'Eastern Time (US & Canada) (UTC-5)'. It has a recurrence of 'Every 1 Day(s)'. The 'Advanced recurrence options' section includes fields for 'Hours' and 'Minutes' (set to 22:05), and a checkbox for 'Specify an end date'. The 'Annotations' section has a '+ New' button. The 'Status' is set to 'Started'.

The screenshot shows the 'Trigger runs' page in the Microsoft Azure Data Factory portal. The left sidebar lists 'Runs', 'Pipeline runs', 'Trigger runs' (which is selected), 'Change Data Capture (previous)', 'Runtimes & sessions', 'Integration runtimes', 'Data flow debug', and 'Notifications'. The main area displays a table of trigger runs. The table has columns: Trigger name, Trigger type, Trigger time, Status, Pipelines, Run, Message, Properties, and Run ID. There are two entries:

Trigger name	Trigger type	Trigger time	Status	Pipelines	Run	Message	Properties	Run ID
CustomersTrigger	Schedule trigger	12/7/2023, 10:12:00 PM	<span style="color: green;">Succeeded</span>	1	Original			085849960176523698077200
RatingsTrigger	Schedule trigger	12/7/2023, 10:04:59 PM	<span style="color: green;">Succeeded</span>	1	Original			085849960218554124792456

# DATA REFRESH-CUSTOMERS

Screenshot of the Azure Data Factory pipeline editor showing the configuration of a trigger.

**Edit trigger**

**Name:** CustomersTrigger

**Description:** (Empty)

**Type:** ScheduleTrigger

**Start date:** 12/7/2023, 10:12:00PM

**Time zone:** Eastern Time (US & Canada) (UTC-5)

This time zone observes daylight savings. Trigger will auto-adjust for one hour difference.

**Recurrence:** Every 1 Day(s)

**Advanced recurrence options:**

**Execute at these times:**

Hours: (Empty)

Minutes: (Empty)

**Schedule execution times:** 22:12

Specify an end date

**Annotations:**

+ New

**Output**

Pipeline run ID: 5aa6d37b-e99f-438e-8f09-bf36cc6f132a

All status

Showing 1 - 1 of 1 items

Activity name	Activity status	Activity type	Run start	Duration
Document-SP	<span style="color: green;">✓ Succeeded</span>	Stored procedure	12/7/2023, 9:48:48 PM	2s

Screenshot of the Microsoft Azure Data Factory Trigger runs page.

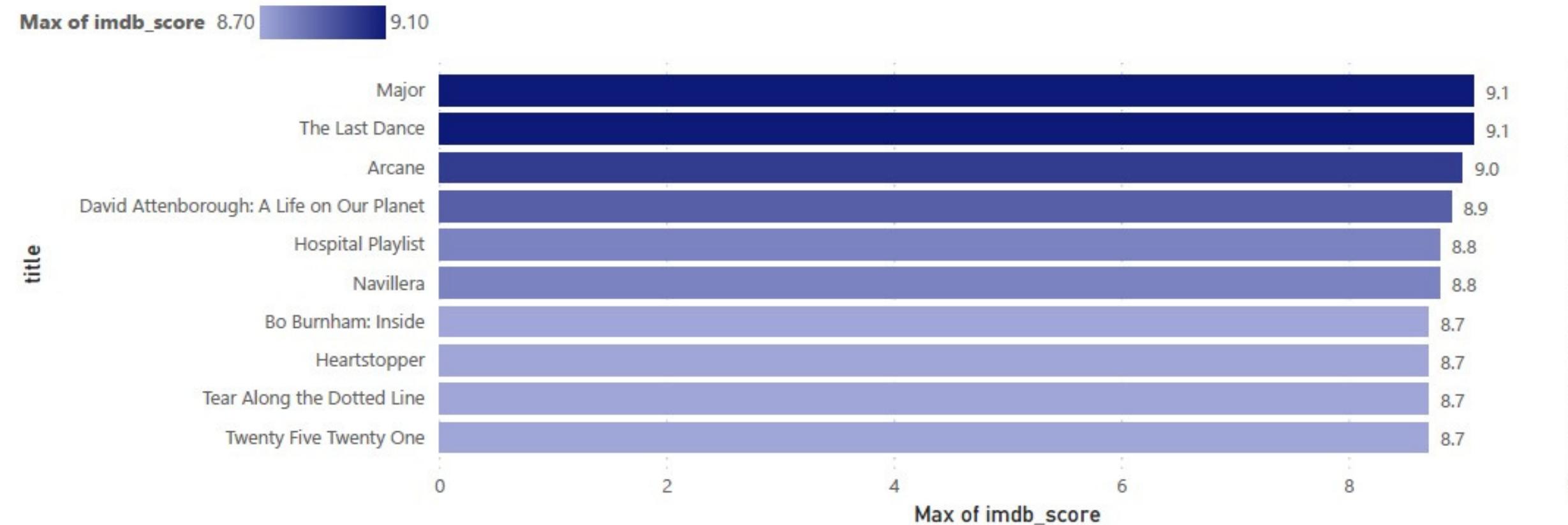
**Trigger runs**

Local time : Last 24 hours | Trigger name : All | Status : All | Runs : Latest runs | Refresh | Edit columns | Export to CSV

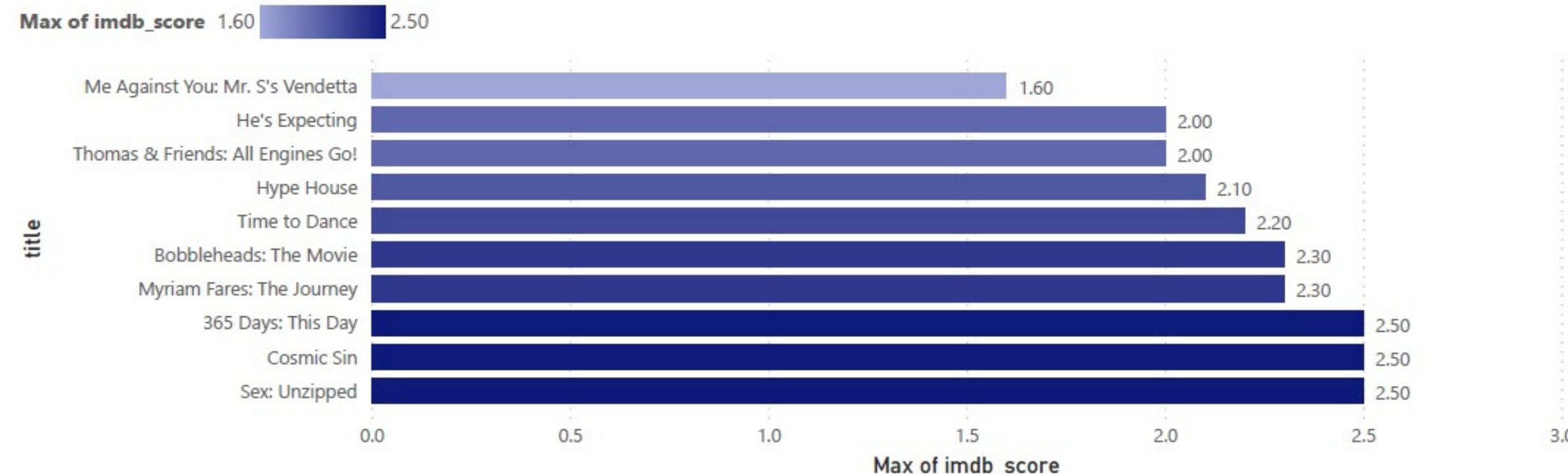
Trigger name	Trigger type	Trigger time	Status	Pipelines	Run	Message	Properties	Run ID
CustomersTrigger	Schedule trigger	12/7/2023, 10:12:00 PM	<span style="color: green;">✓ Succeeded</span>	1	Original			08584996017652369807720
RatingsTrigger	Schedule trigger	12/7/2023, 10:04:59 PM	<span style="color: green;">✓ Succeeded</span>	1	Original			08584996021855412479245

# VISUALIZATIONS

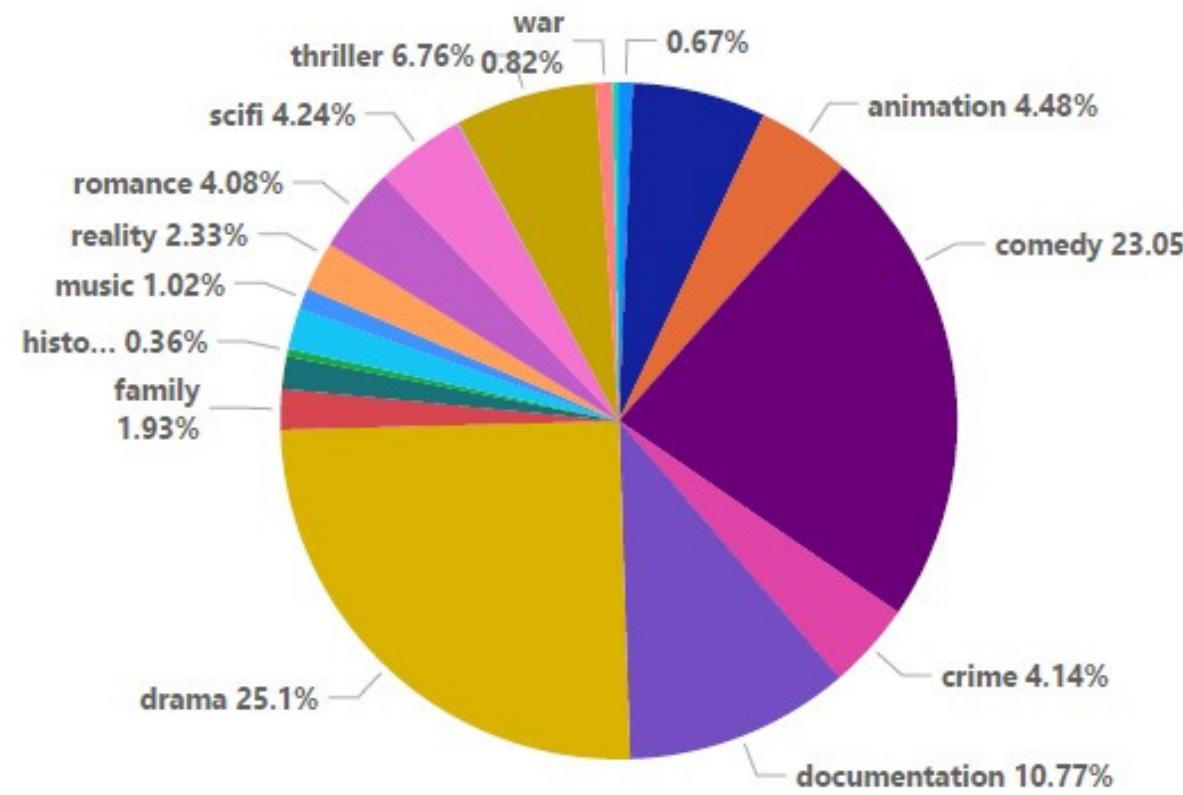
## Top 10 Highest Rated Movies watched in 2022



## Top 10 Lowest Rated Movies watched in 2022



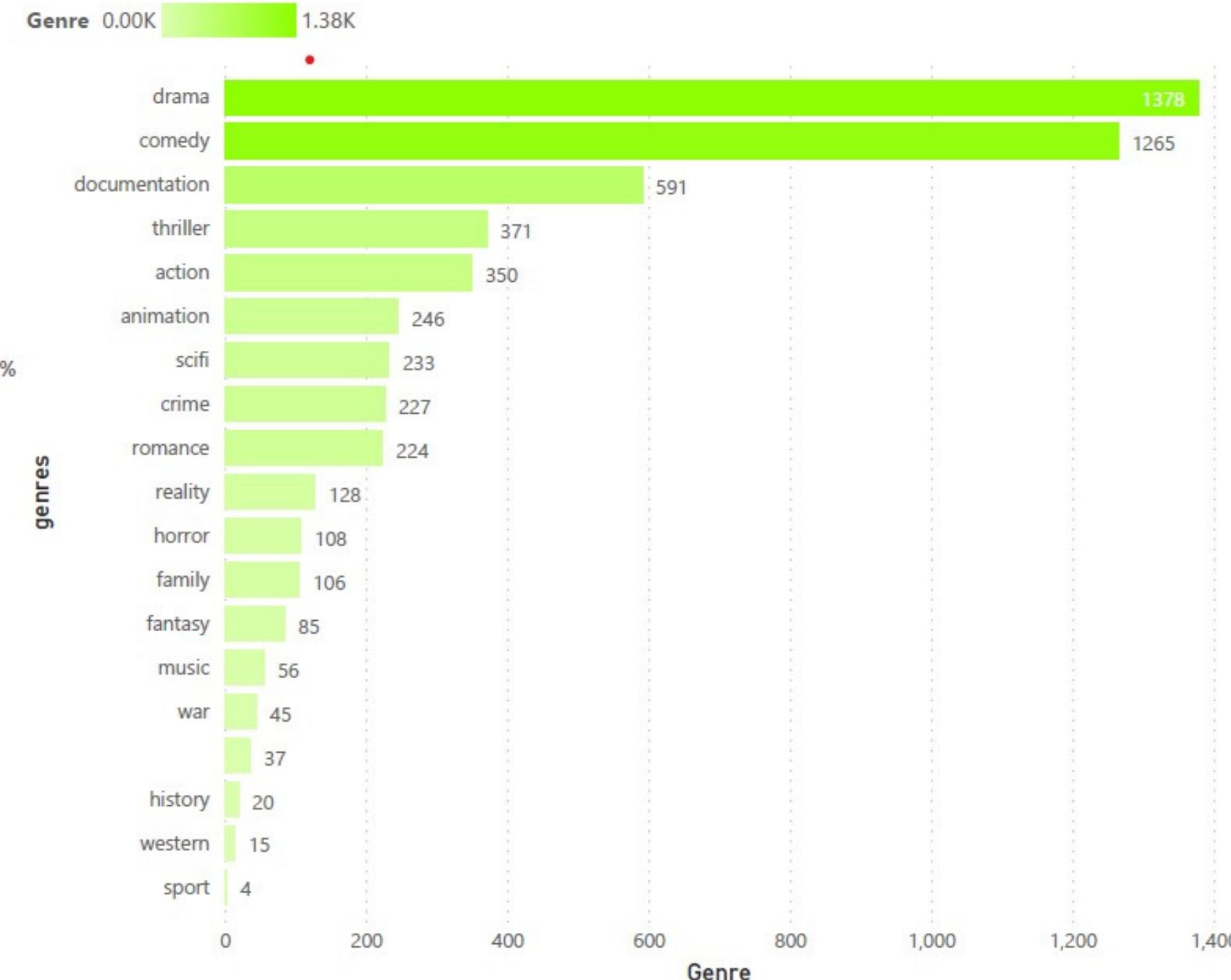
### Top Rated Genre's



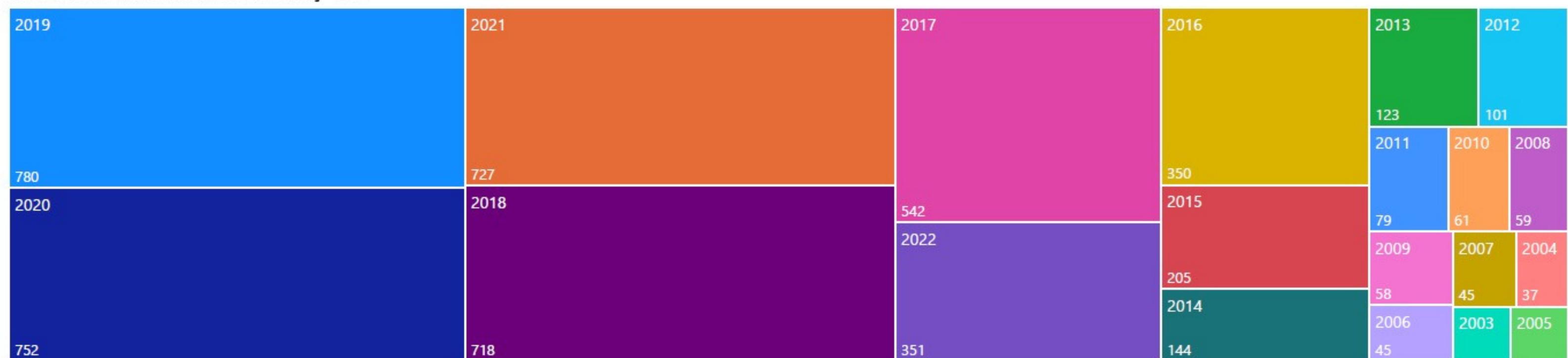
19

Count of genres

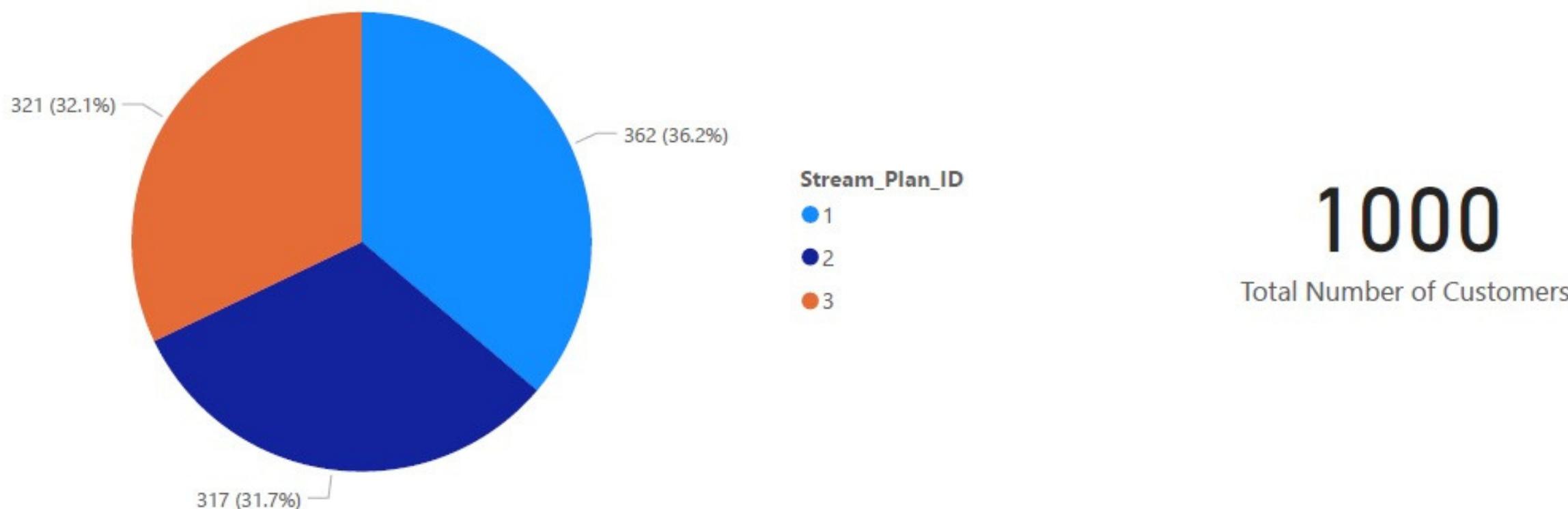
### Number of Movies by Genre



### Number of Movies Released Every Year

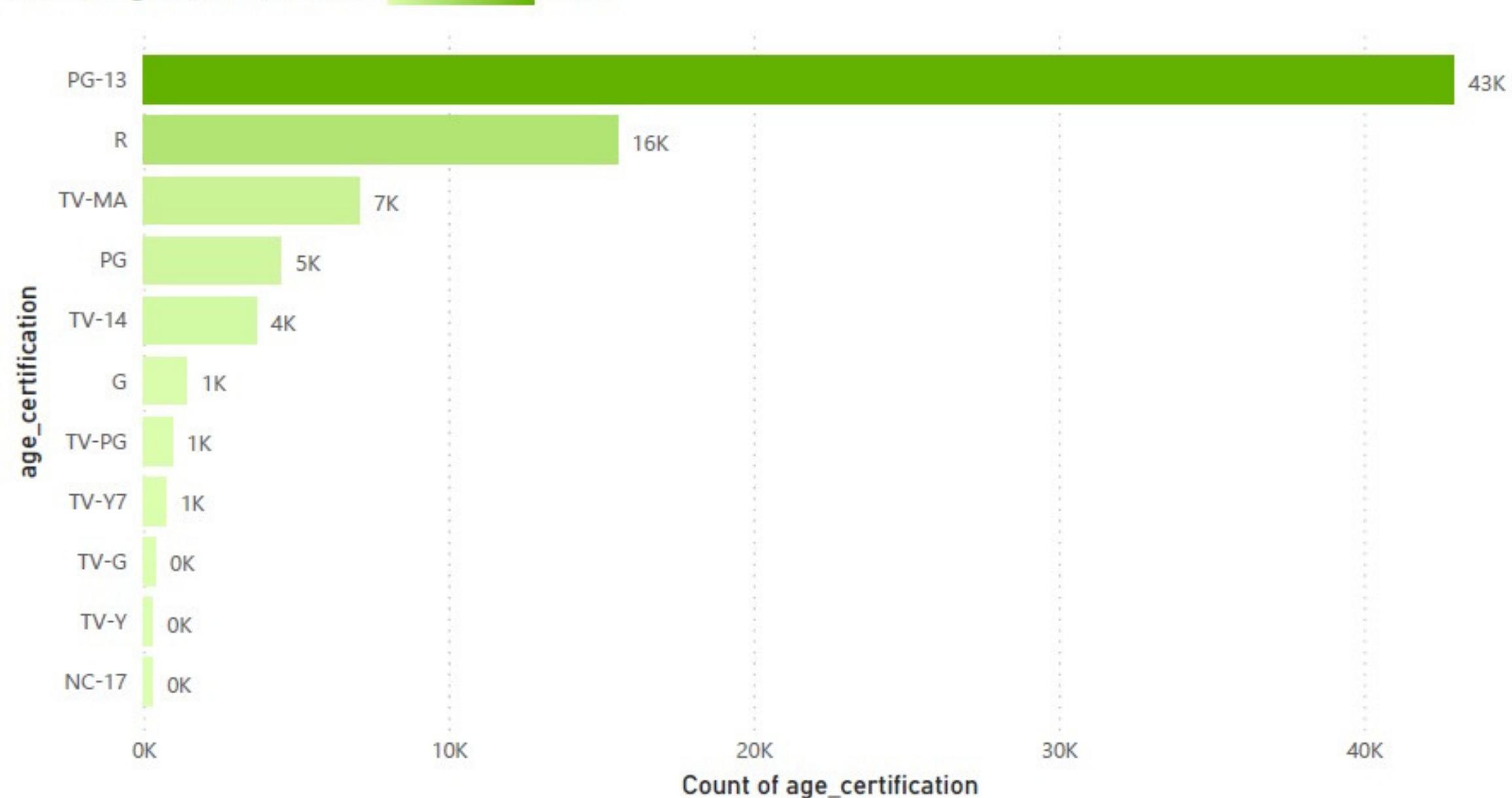


### Total Percentage of Customers with Each Stream Plan



### Number of Movies with Age Certification

Count of age\_certification 0.29K 42.95K



### Top 3 Actors with Highest Movie Count

Count of id first\_name last\_name

25	Boman	Irani
25	Kareena	Kapoor Khan
23	Shah	Rukh Khan

**77.80K**

Total Number of Actors

**5489**

Total Number of Movies

# FUTURE PLANS



Implement more sophisticated data analytics methods to gain deeper insights from the streaming service data. This could involve using machine learning models to predict user behavior or preferences.

## ADVANCED DATA ANALYTICS INTEGRATION



Develop more complex and interactive data visualizations to better understand user trends and patterns. This can help in making data-driven decisions more effectively..

## ENHANCED DATA VISUALIZATION



Focus on scaling the database and data processing pipelines to handle an increased volume of data efficiently. This includes optimizing existing processes for better performance.

## SCALABILITY AND PERFORMANCE OPTIMIZATION

# **THANK YOU!**

# **QUESTIONS?**

