Big Data Management and Analysis

Project Report
On
Data Visualisation using MongoDB Atlas Charts

Submitted to:
Prof (Dr.) Amarnath Mitra

Submitted By:
Preksha Verma (055032)
Suvra Datta Banik (055049)

PGDM Big Data Analytics-05 | Section K

1. Project Introduction

The M.Flix Database System is designed to efficiently manage movie-related data, including film details, ratings, genres, and award statistics. To ensure the reliability of the system, the database must maintain data integrity, proper indexing, and performance efficiency under high query loads. This report documents a series of validation and stress tests conducted to verify compliance with data structuring principles, enforce relational integrity where applicable, and evaluate system performance under large-scale data operations. These tests help confirm that the database can handle real-world query demands effectively.

2. Description of Data

2.1. Shape of dataset:

Number of Variables: 135

Number of Observations: 21,349

2.2.Data Structure & Variable Types

2.2.1. Text Variables (Categorical)

2.2.1.1. Movie Metadata:

a) _id (Unique Movie ID)
b) plot (Movie Synopsis)
c) genres[0], genres[1], genres[2] (Movie Genres)
d) cast[0], cast[1], cast[2], ... (Actors)
e) writers[0], writers[1], ... (Writers)
f) directors[0], directors[1], ... (Directors)
g) language (Primary Language)
h) country (Production Country)
i) awards (Awards Won/Nominated)
j) production (Production Company)

2.2.1.2. Rotten Tomatoes Metadata:

a) tomatoes.consensus (Critics' Summary)
b) tomatoes.website (Official Movie Website)
c) tomatoes.production (Production Studio)

2.2.2. Numerical Variables

2.2.2.1. Movie Metrics:

a) runtime (Duration in minutes)
b) metacritic (Metacritic Score)
c) tomatoes.boxOffice (Box Office Revenue)

2.2.3. Time Variables

a) tomatoes.dvd (DVD Release Date)
b) releaseDate (Theatrical Release Date)

3. Problem Objectives

A well-structured movie database must meet the following criteria:

3.1. Ensure data consistency through unique identifiers and reference constraints where necessary.
3.2. Maintain data accuracy to prevent duplicate or inconsistent records.
3.3. Optimize indexing strategies to enhance search efficiency.
3.4. Handle concurrent queries efficiently without performance bottlenecks.
3.5. Support complex queries, such as filtering by multiple attributes (genres, ratings, release years).
3.6. Maintain performance even under high data loads and bulk update operations.

3.7. To address these challenges, a series of validation and stress tests were conducted, evaluating the M.Flix database's ability to process large datasets while maintaining accuracy, efficiency, and scalability.

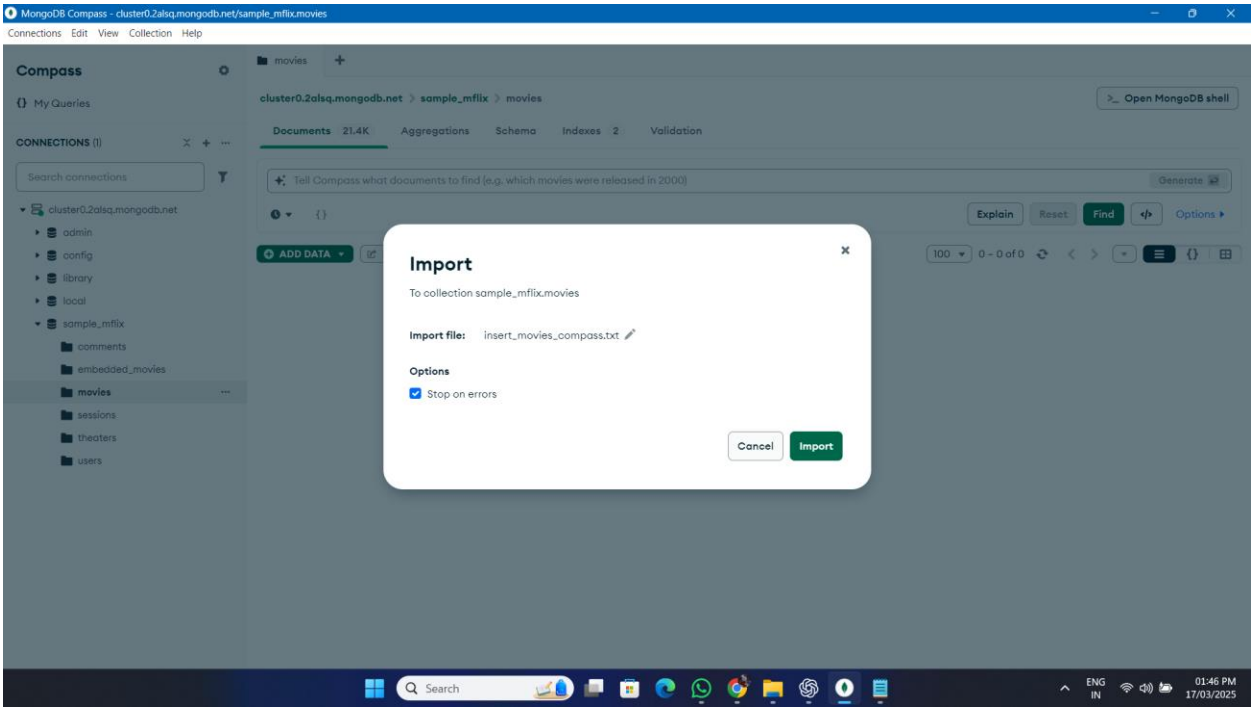
4. Tools Used

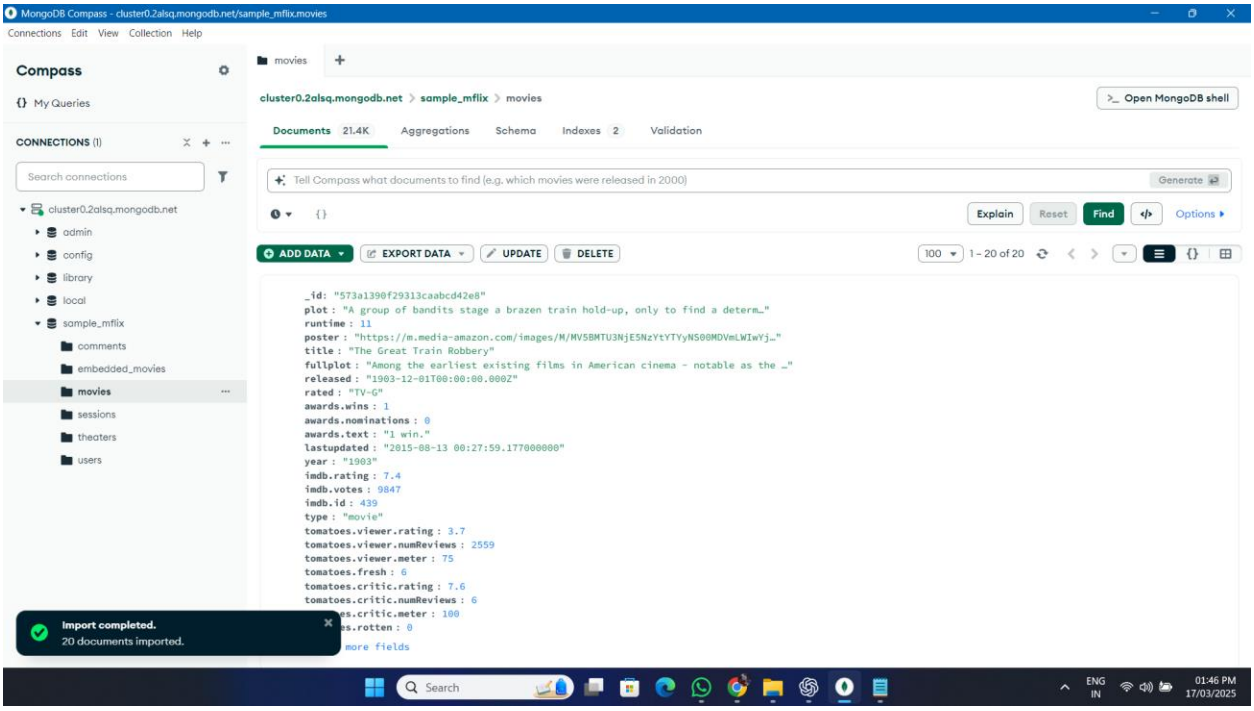
The following tools and technologies were used for database testing:

4.1. **MongoDB** – NoSQL database system used for managing and querying movie data.
4.2. **MongoDB Query Language (MQL)** – Used to perform validation checks, retrieve data, and test indexing strategies.
4.3. **MongoDB Compass** – A tool used for analyzing queries and monitoring performance metrics.
4.4. **MongoDB Atlas charts**- A graphical tool used for creating charts for data visualisation.
4.5. **Command Line Interface (CLI)** – Used to execute bulk data operations and stress testing.

5. Operations Performed

5.1. Import

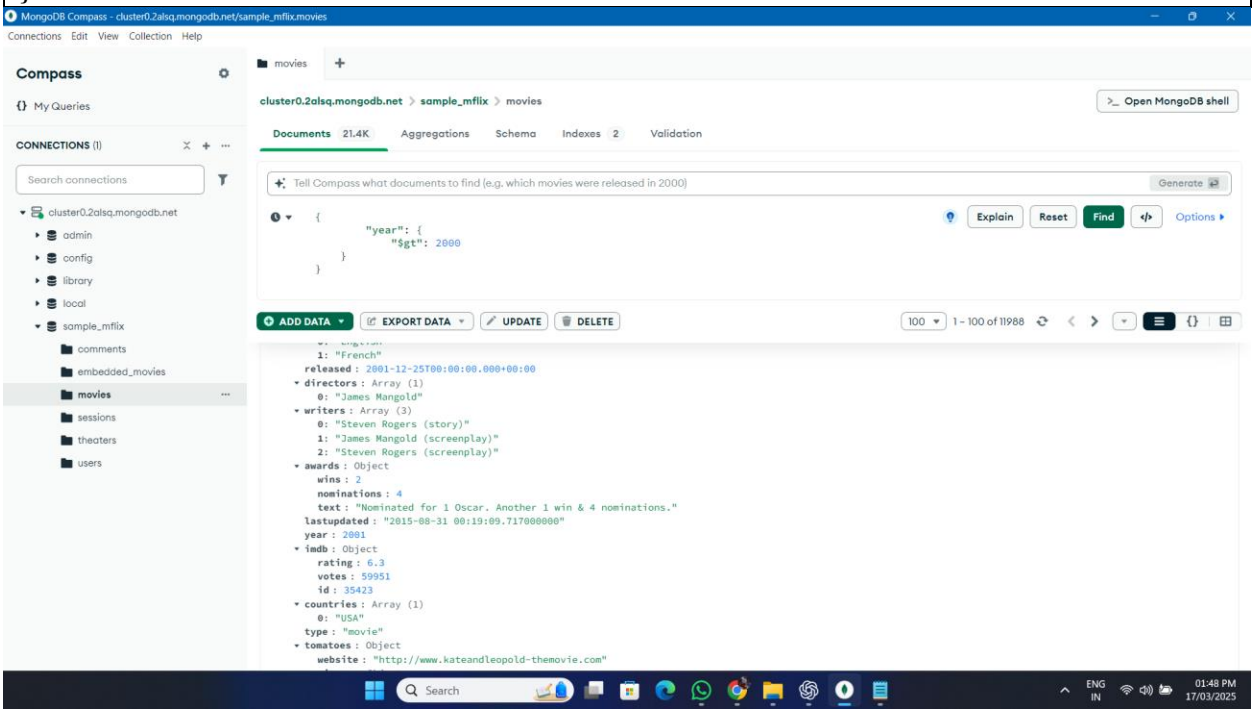




5.2. Retrieve

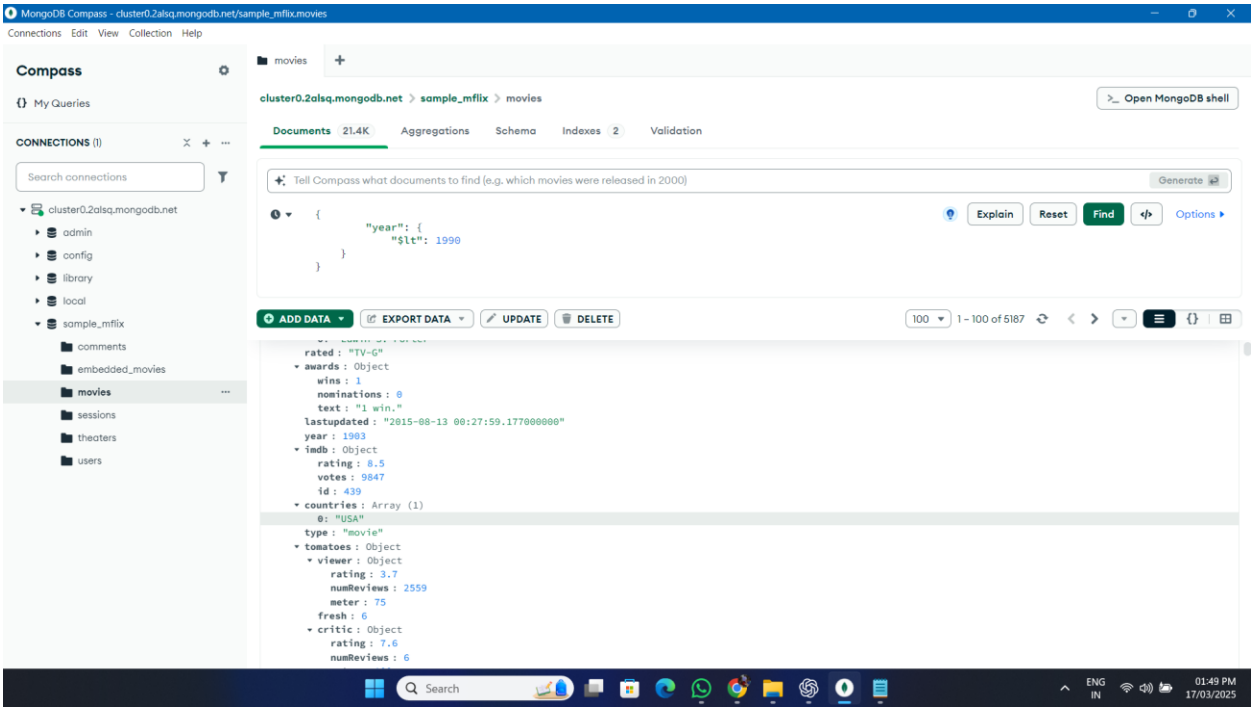
5.2.1. Retrival of data for year 2000

```
{
      "year": {
         "$gt": 2000
   }
}
```



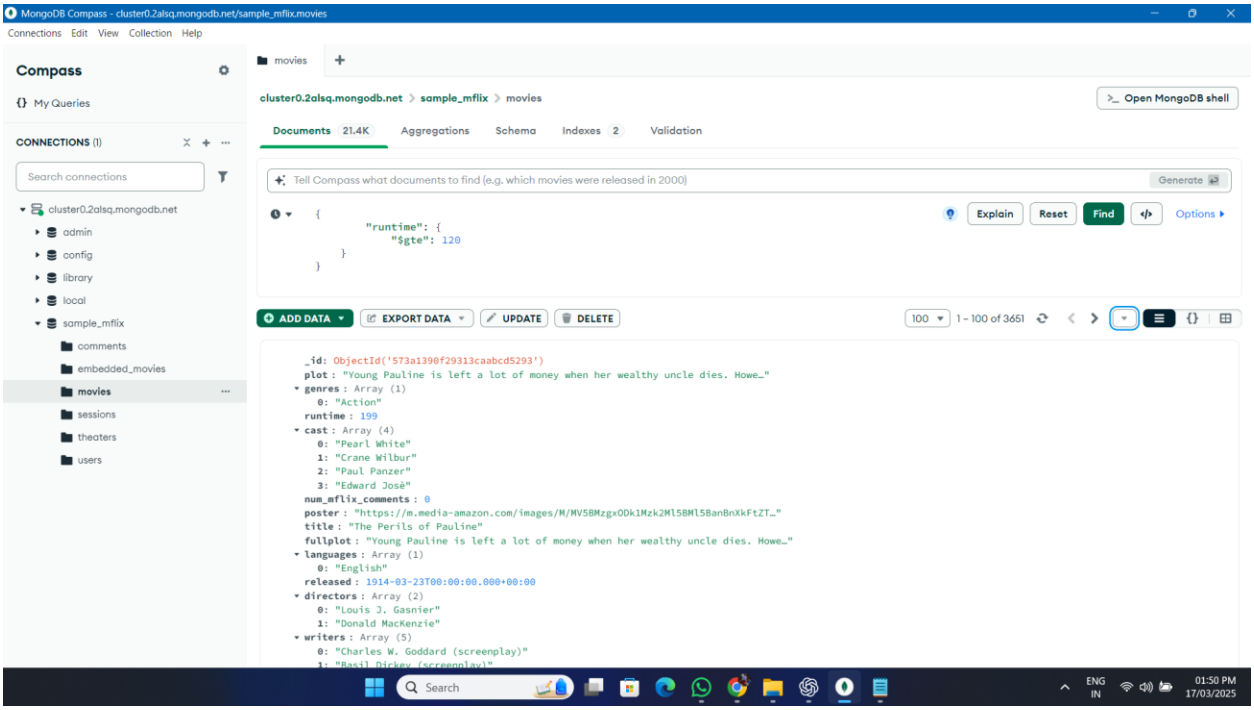
5.2.2. Retrival of data for year 1990

```
{
      "year": {
         "$lt": 1990
   }
}
```



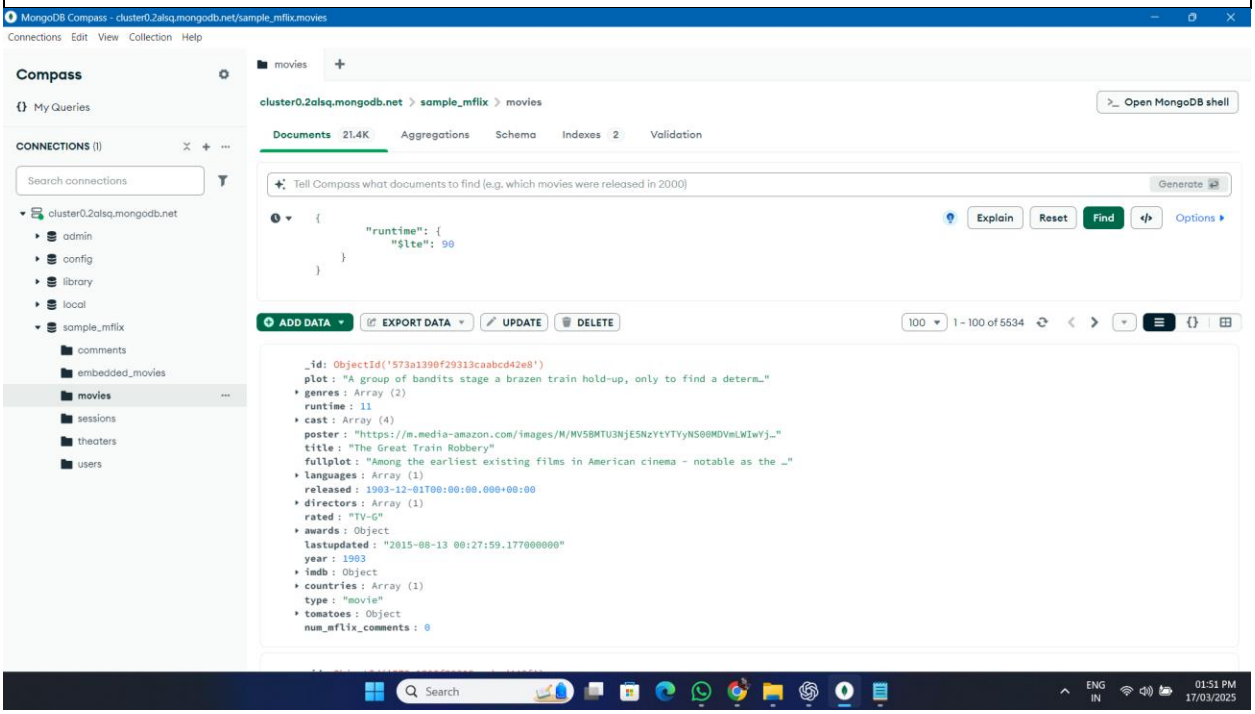
5.2.3. Retrieval of data for runtime 120 minutes

```
{
      "runtime": {
         "$gte": 120
   }
}
```



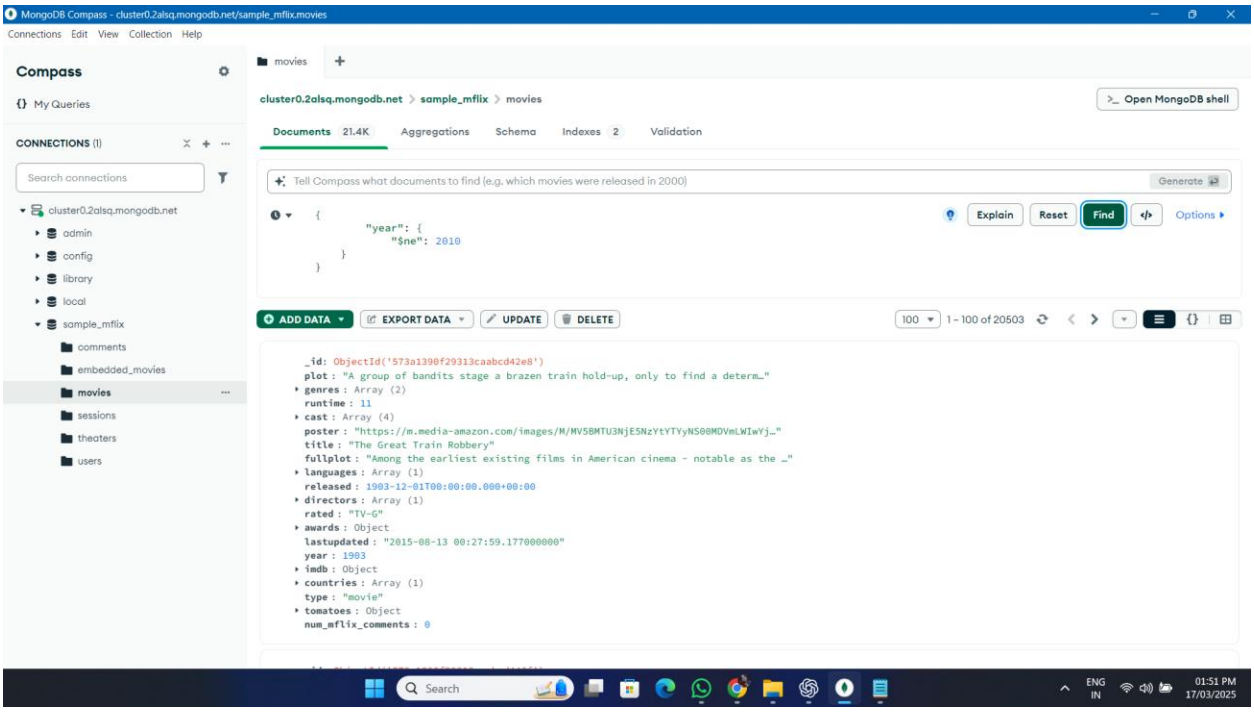
5.2.4. Retrieval of data for runtime 90 minutes

```
{
      "runtime": {
         "$lte": 90
   }
}
```



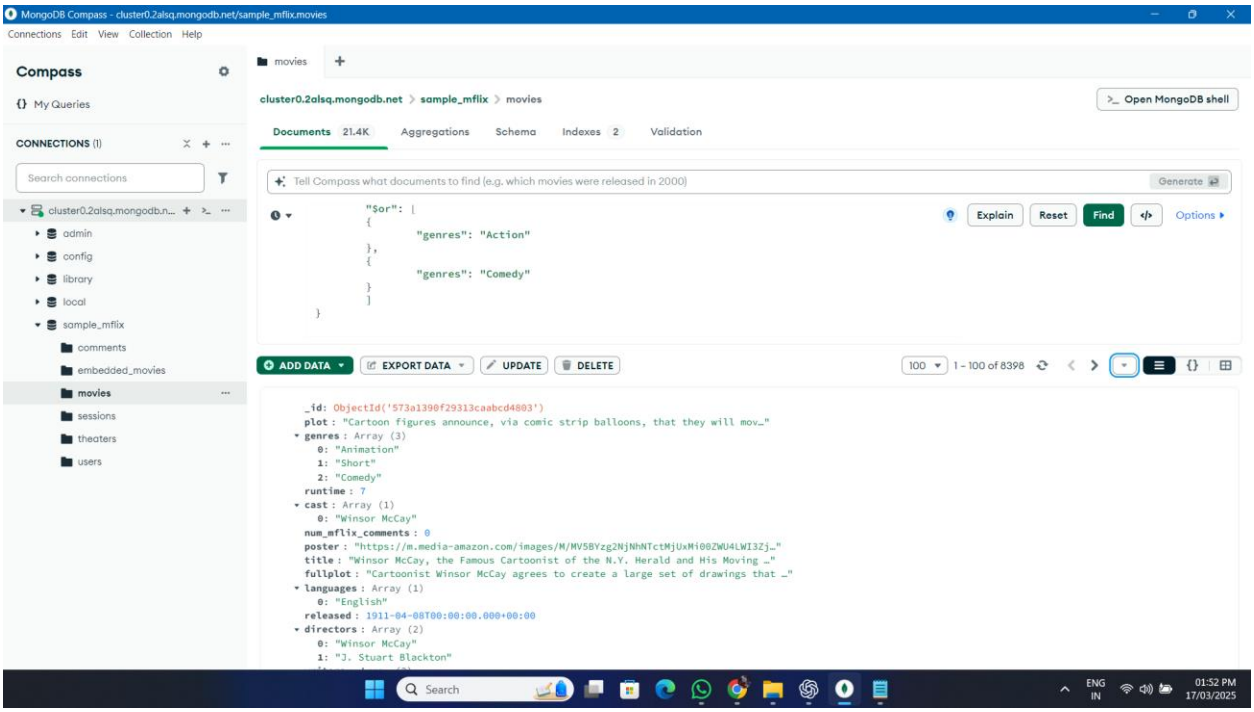
5.2.5. Retrieval of data for runtime 120 minutes

```
{
      "year": {
         "$ne": 2010
   }
}
```



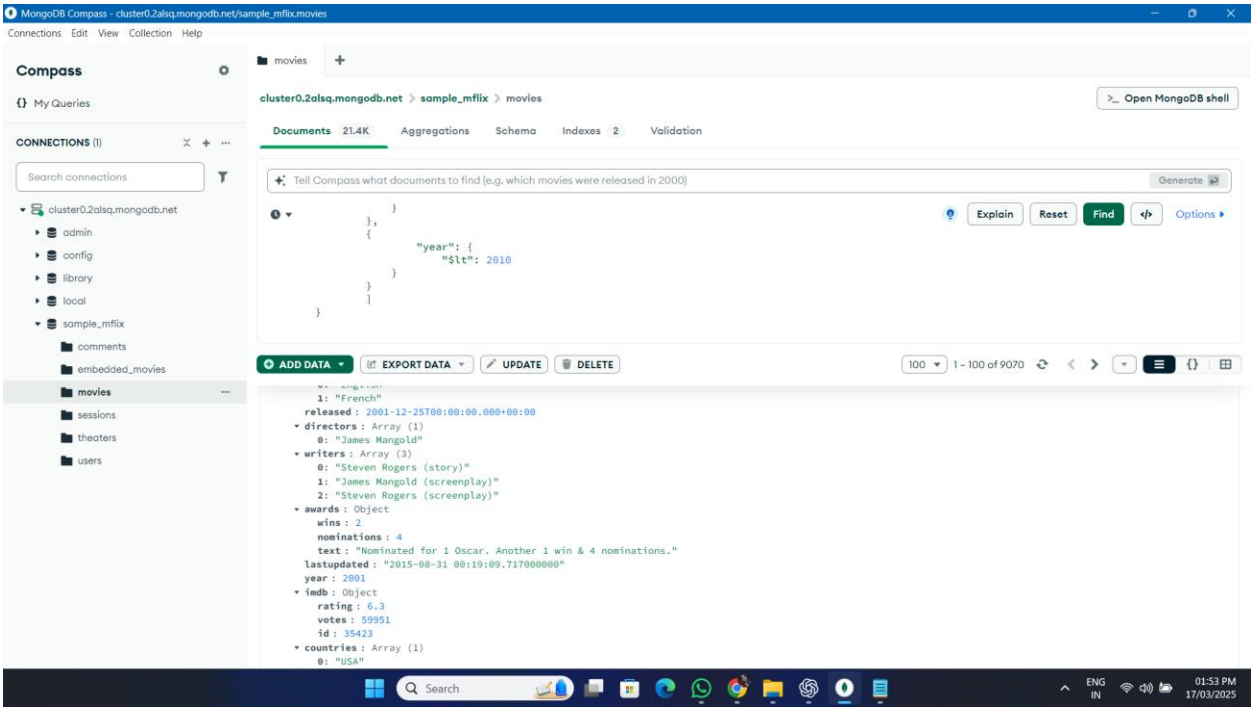
5.2.6. Retrieval of data for genre action

```
{
	"$or": [
	{
		"genres": "Action"
	},
	{
		"genres": "Comedy"
	}
	]
}
```



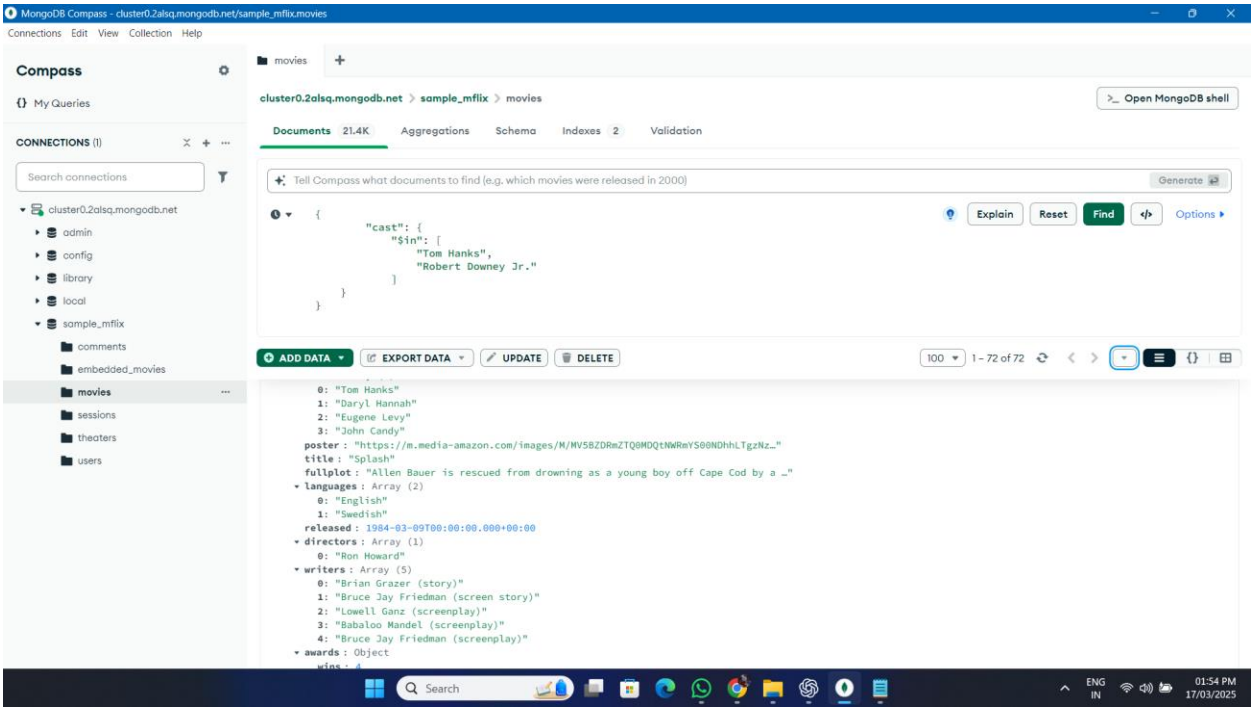
5.2.7. Retrieval of data for duration between year 1995 and 2010

```
{
      "$and": [
      {
            "year": {
               "$gt": 1995
         }
      },
      {
            "year": {
               "$lt": 2010
         }
      }
      ]
}
```



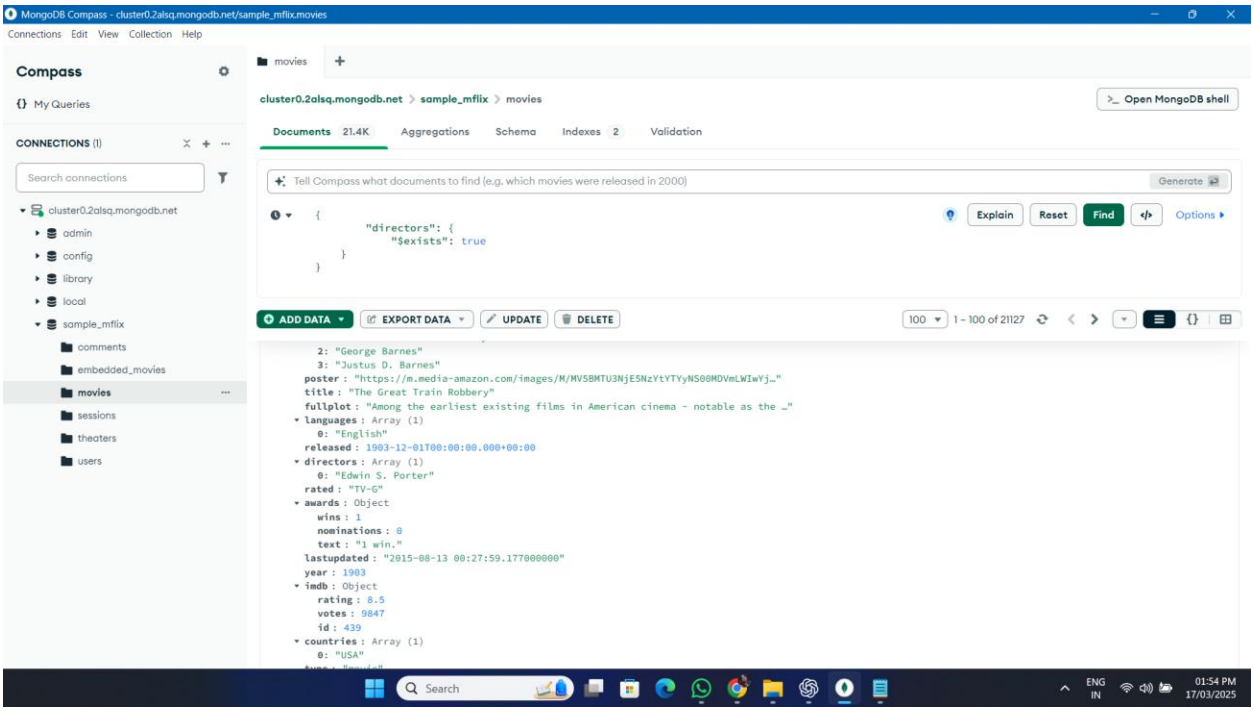
5.2.8. Retrieval of data where Actors are Tom Hanks and Robert Downey Jr.

```
{
      "cast": {
         "$in": [
            "Tom Hanks",
            "Robert Downey Jr."
         ]
   }
}
```



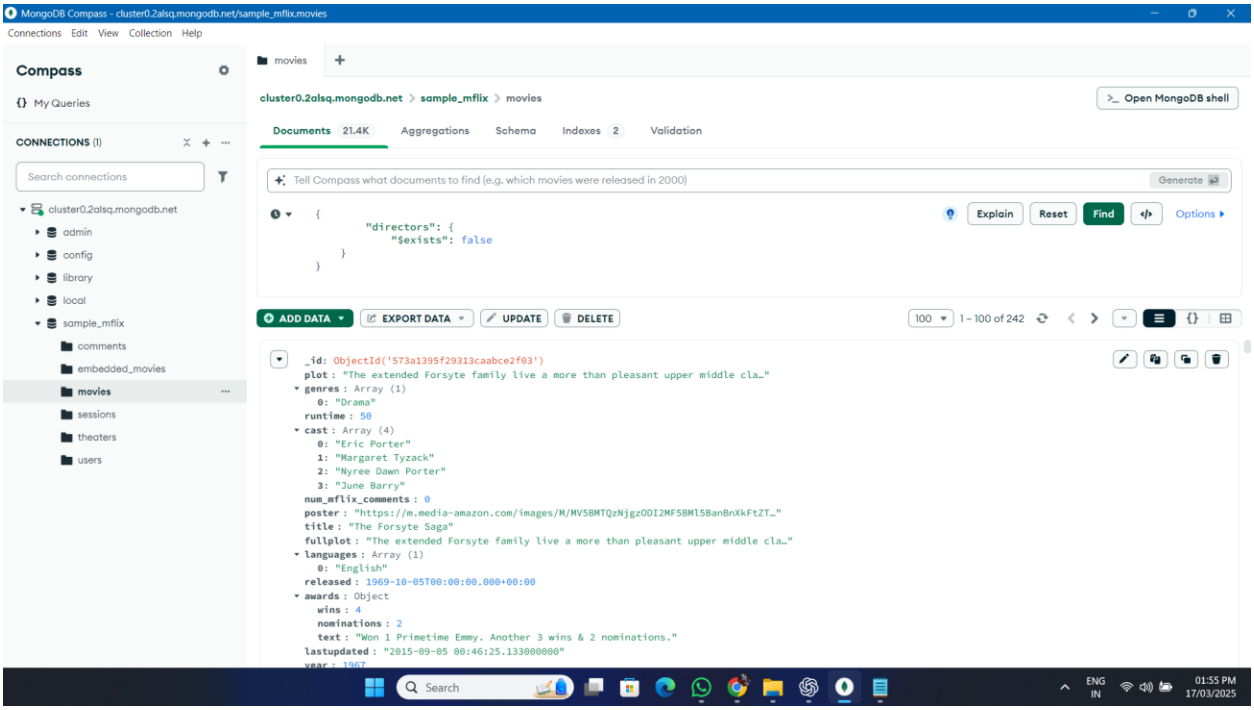
5.2.9. Retrieval of data where information about the director exists

```
{
      "directors": {
         "$exists": true
   }
}
```



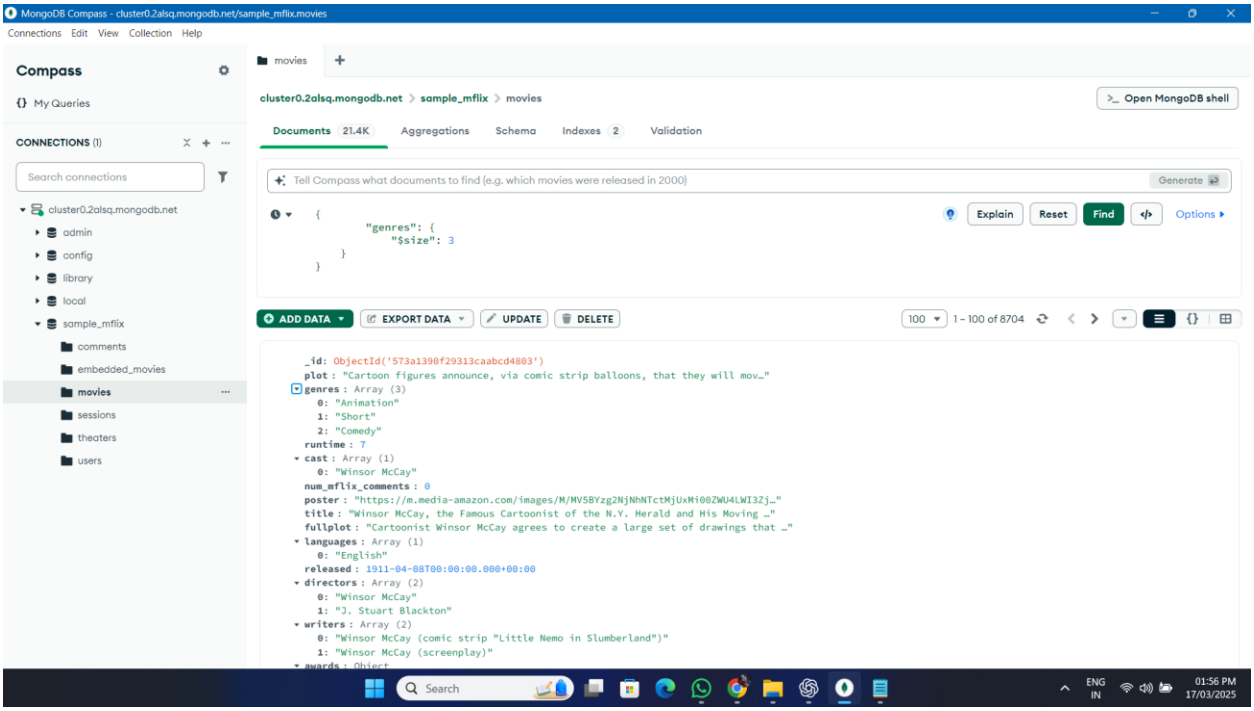
5.2.10. Retrieval of data where information about the director does not exist

```
{
      "directors": {
         "$exists": false
   }
}
```



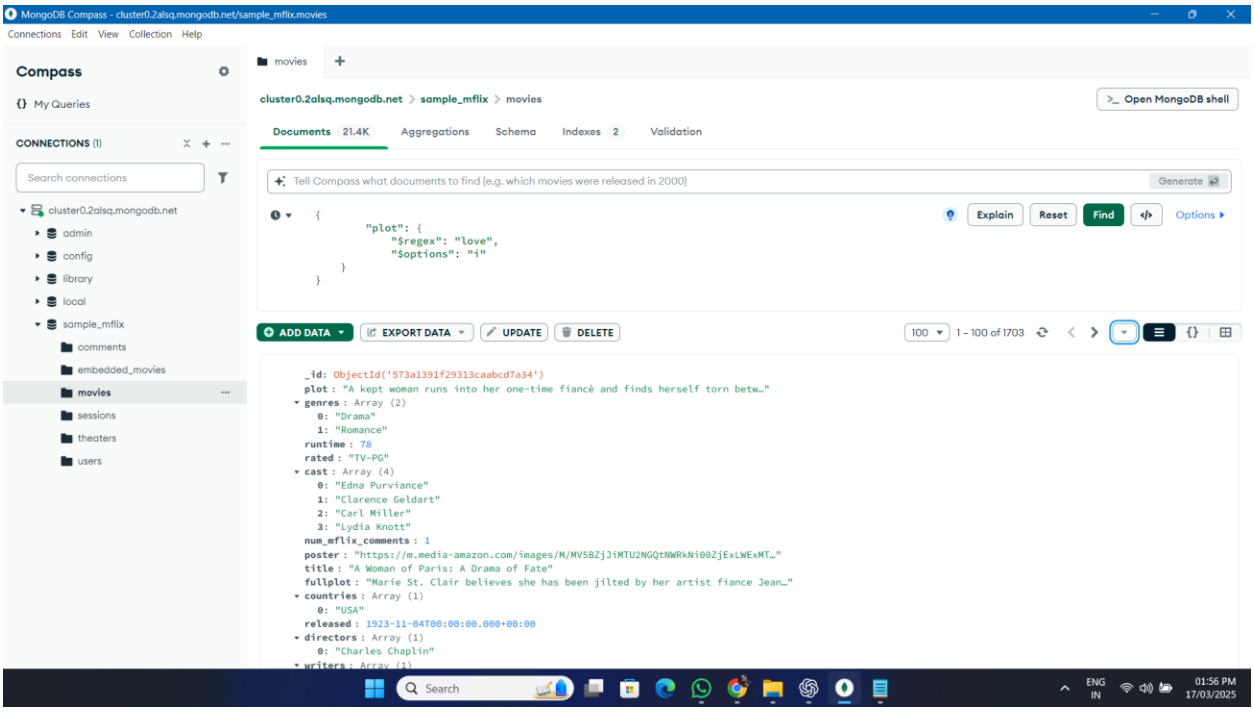
5.2.11. Retrieval of data where genre size is 3.

```
{
      "genres": {
         "$size": 3
   }
}
```



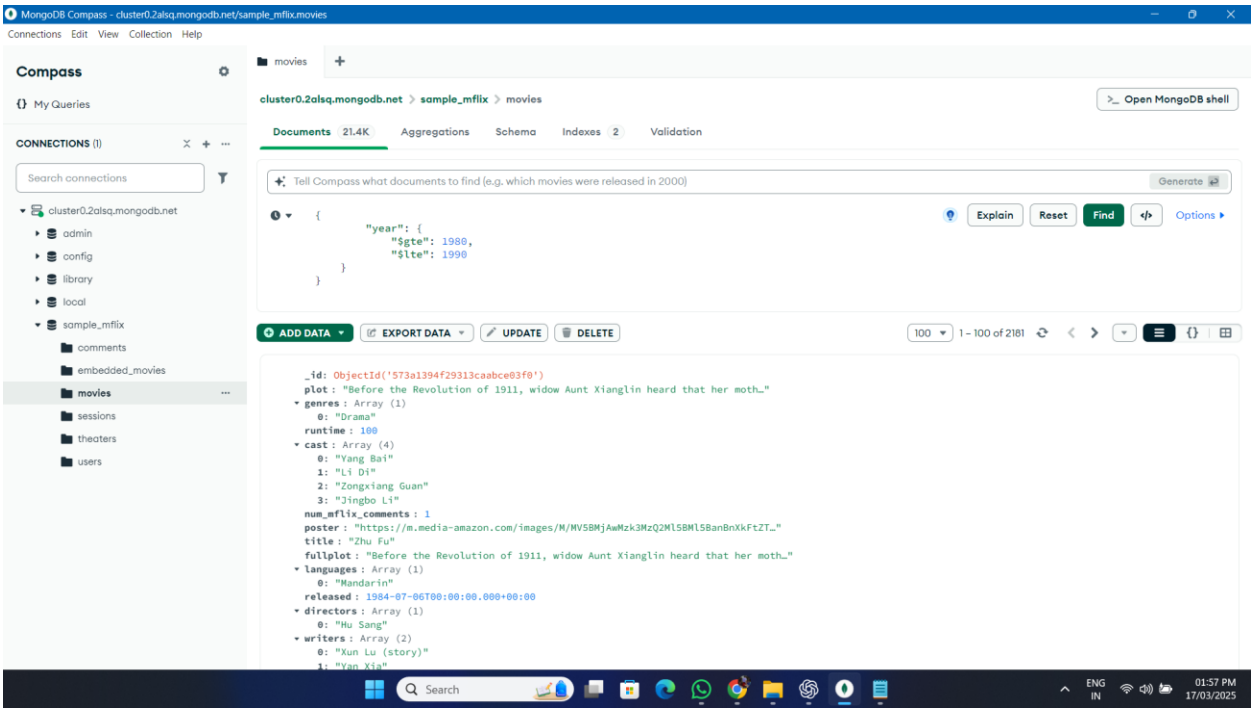
5.2.12.

```
{
      "plot": {
         "$regex": "love",
         "$options": "i"
   }
}
```



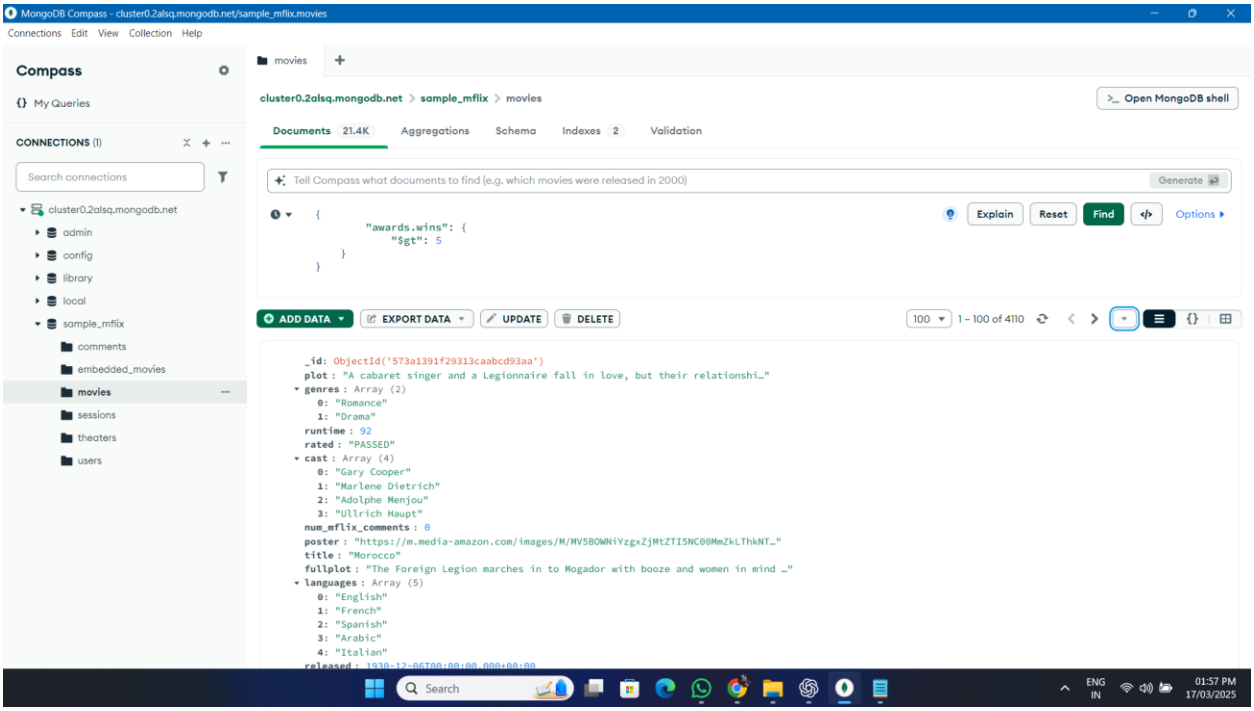
5.2.13.

```
{
      "year": {
         "$gte": 1980,
         "$lte": 1990
   }
}
```



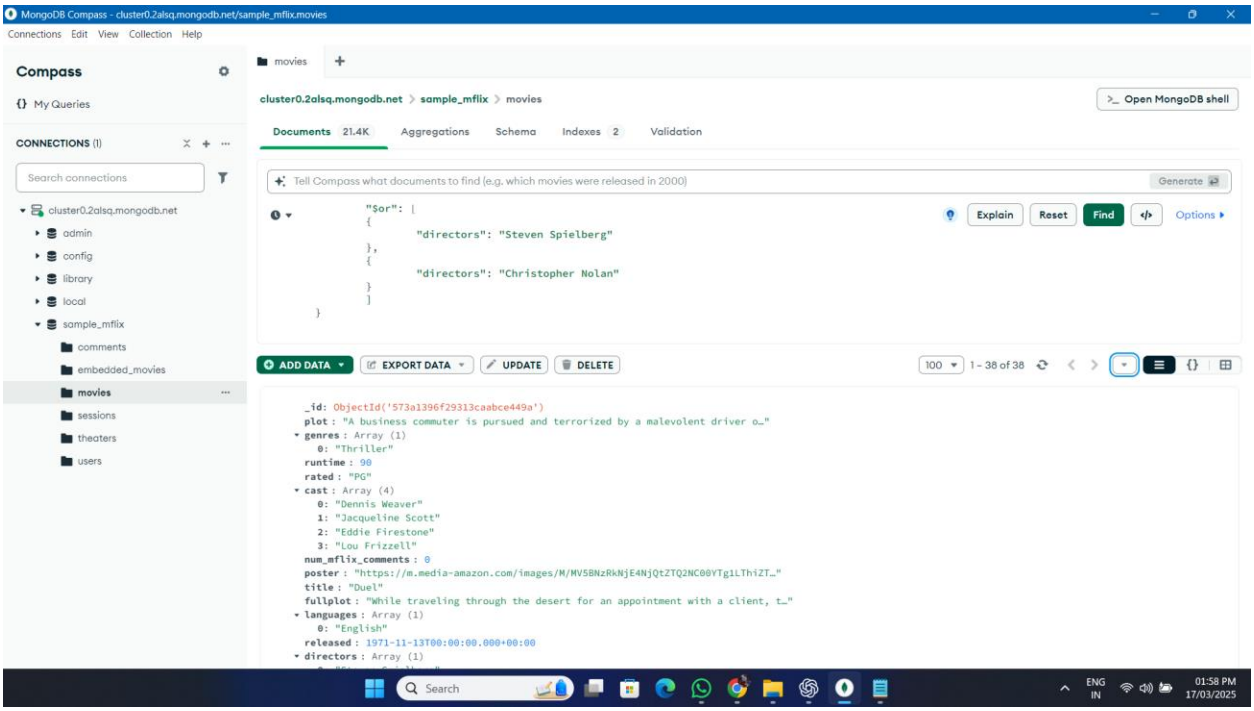
5.2.14.

```
{
      "awards.wins": {
         "$gt": 5
   }
}
```



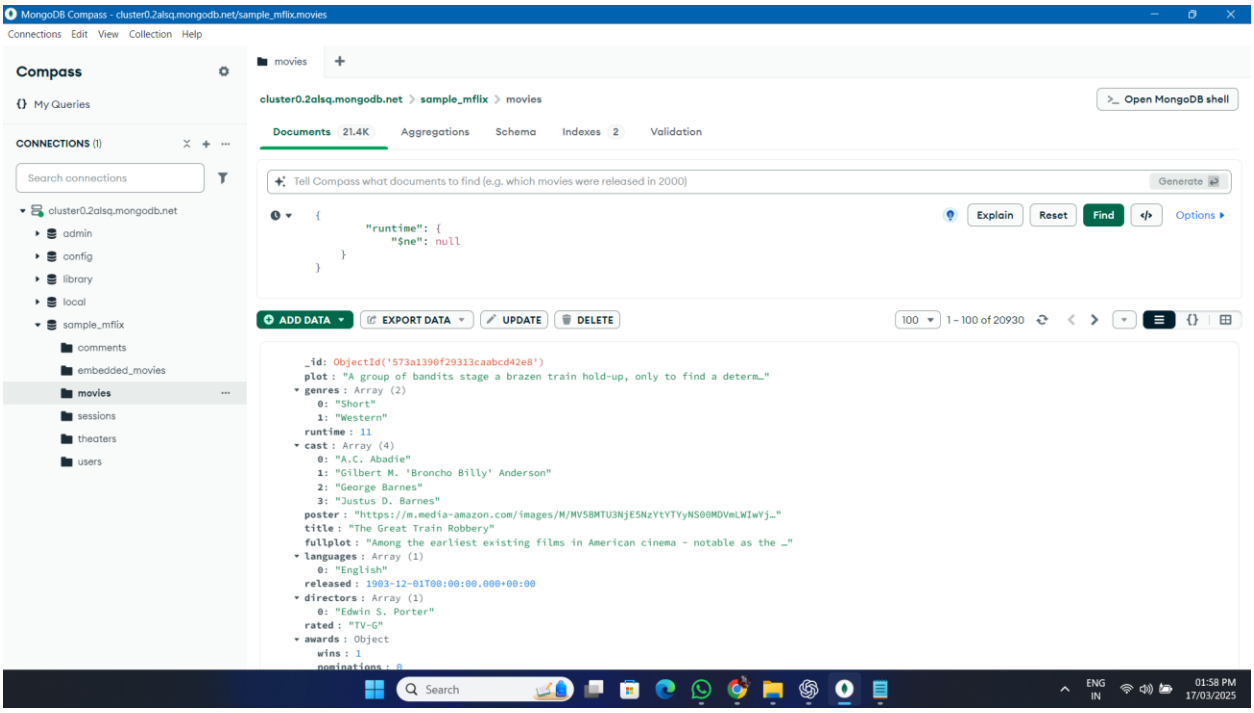
5.2.15.

```
{
	"$or": [
	{
		"directors": "Steven Spielberg"
	},
	{
		"directors": "Christopher Nolan"
	}
	]
}
```



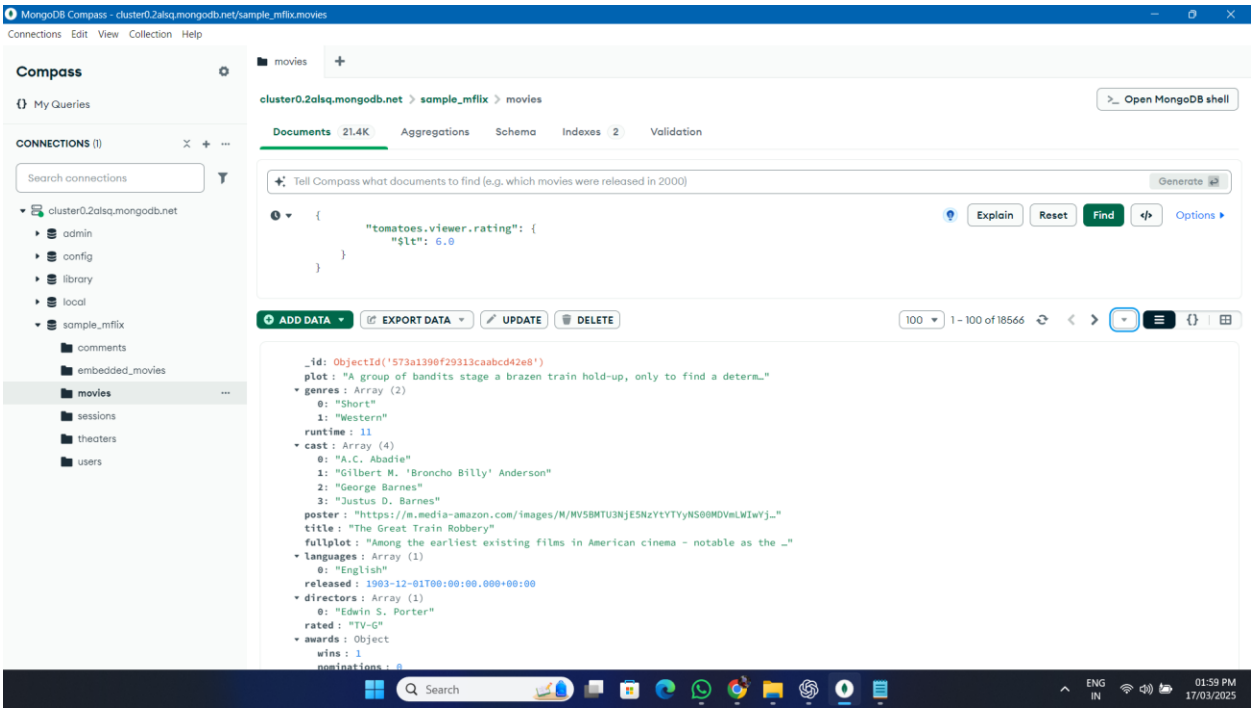
5.2.16.

```
{
      "runtime": {
         "$ne": null
   }
}
```



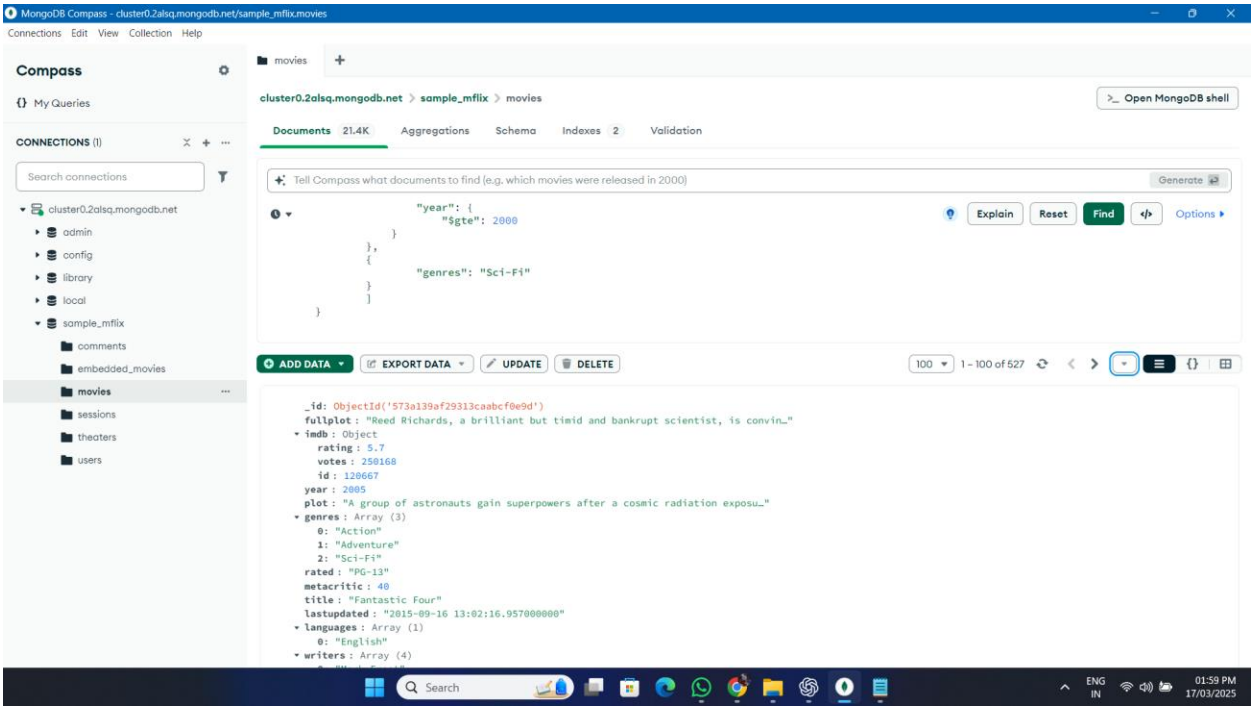
5.2.17.

```
{
        "tomatoes.viewer.rating": {
            "$lt": 6.0
    }
}
```



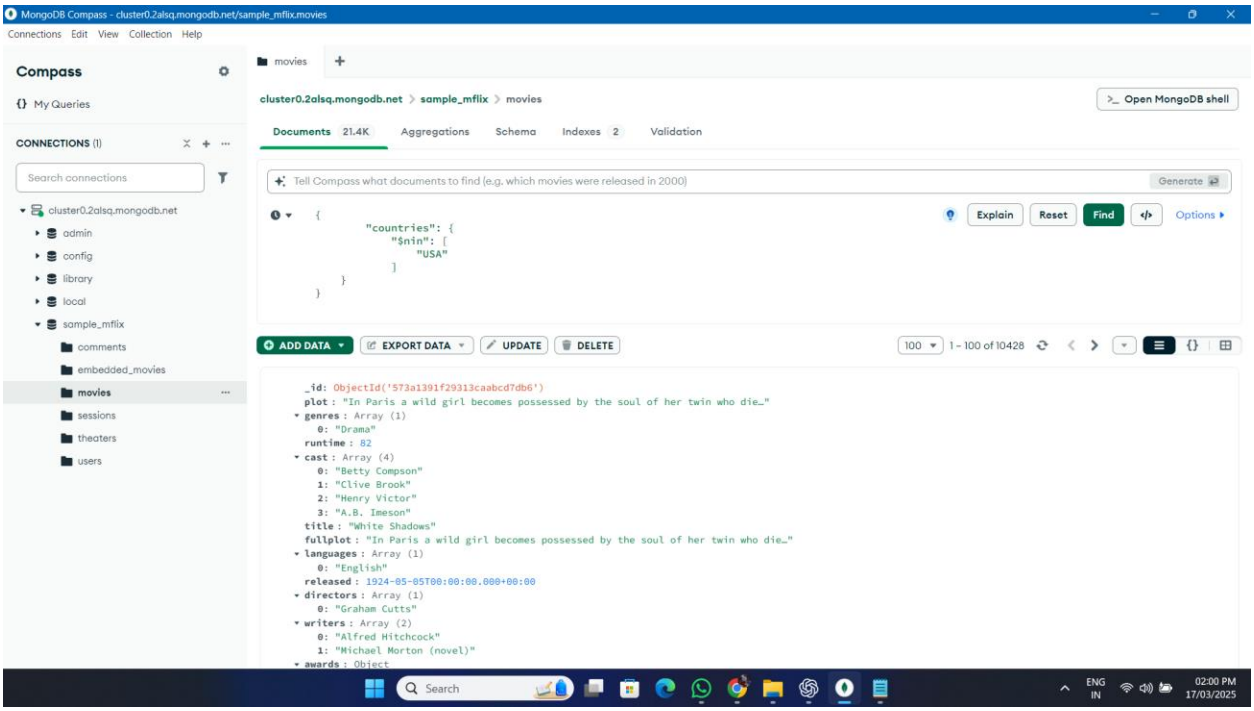
5.2.18.

```
{
      "$and": [
      {
            "year": {
               "$gte": 2000
         }
      },
      {
            "genres": "Sci-Fi"
      }
      ]
}
```



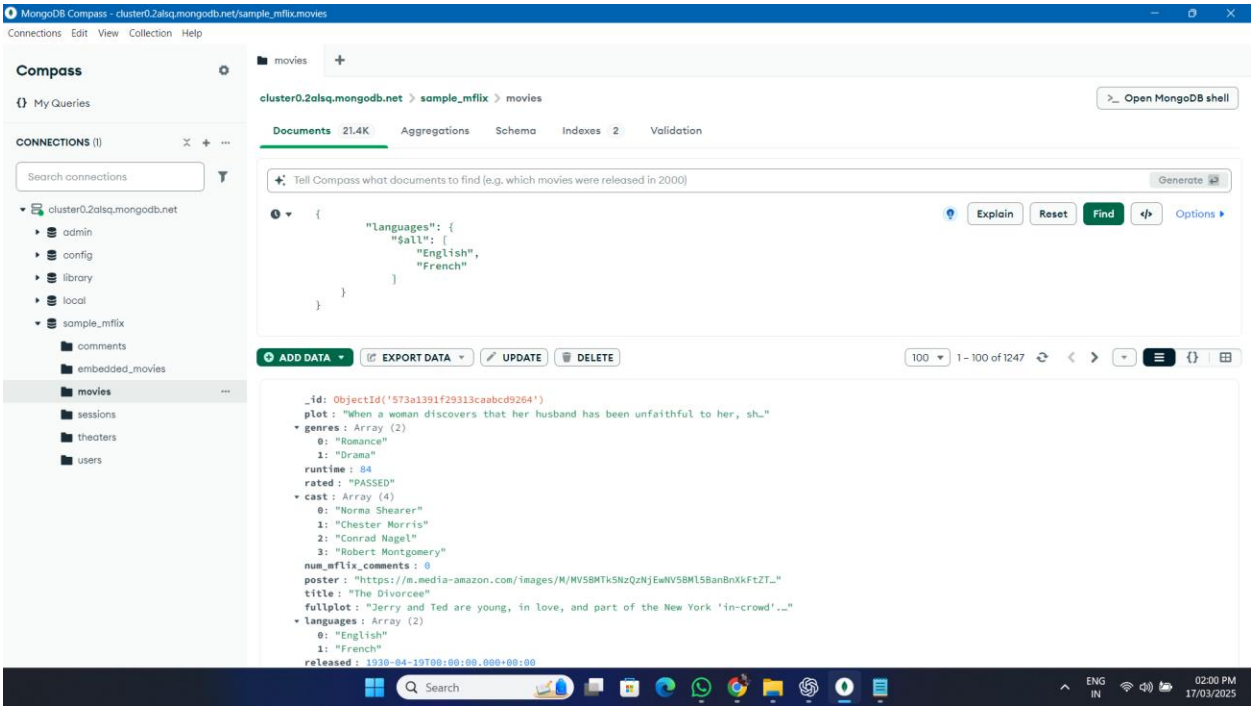
5.2.19.

```
{
      "countries": {
         "$nin": [
            "USA"
         ]
   }
}
```



5.2.20.

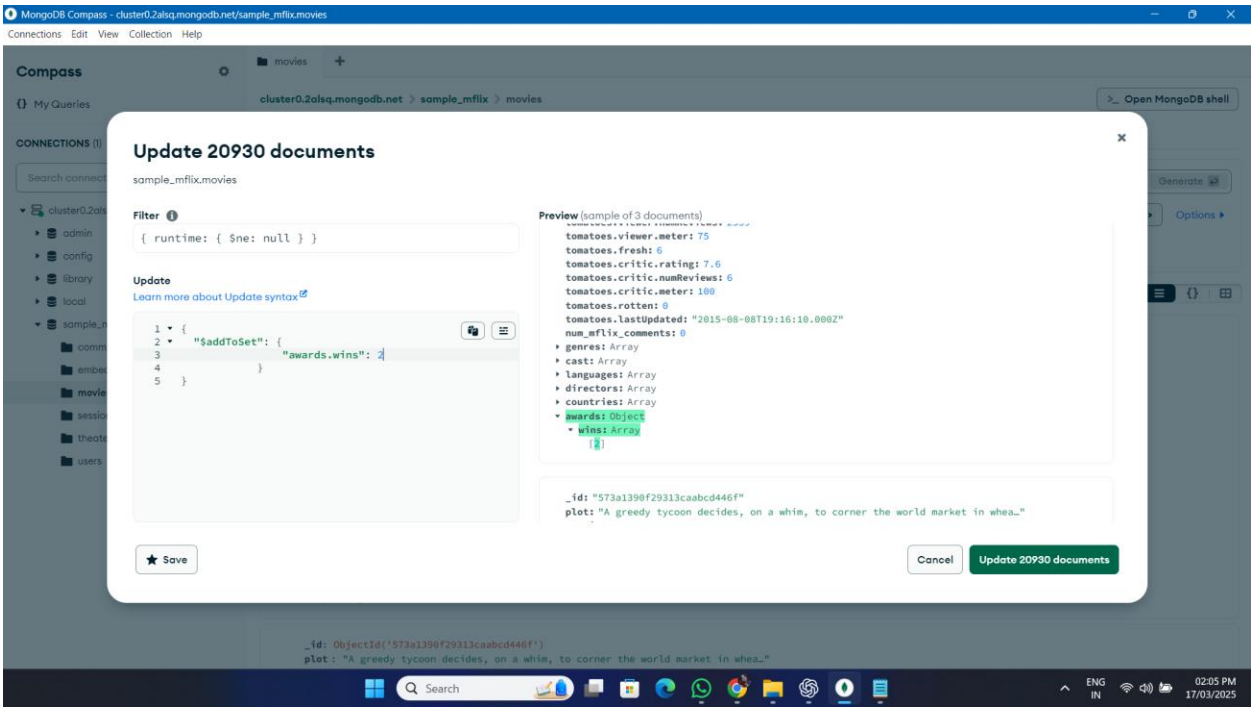
```
{
      "languages": {
         "$all": [
            "English",
            "French"
         ]
   }
}
```



5.3. Update

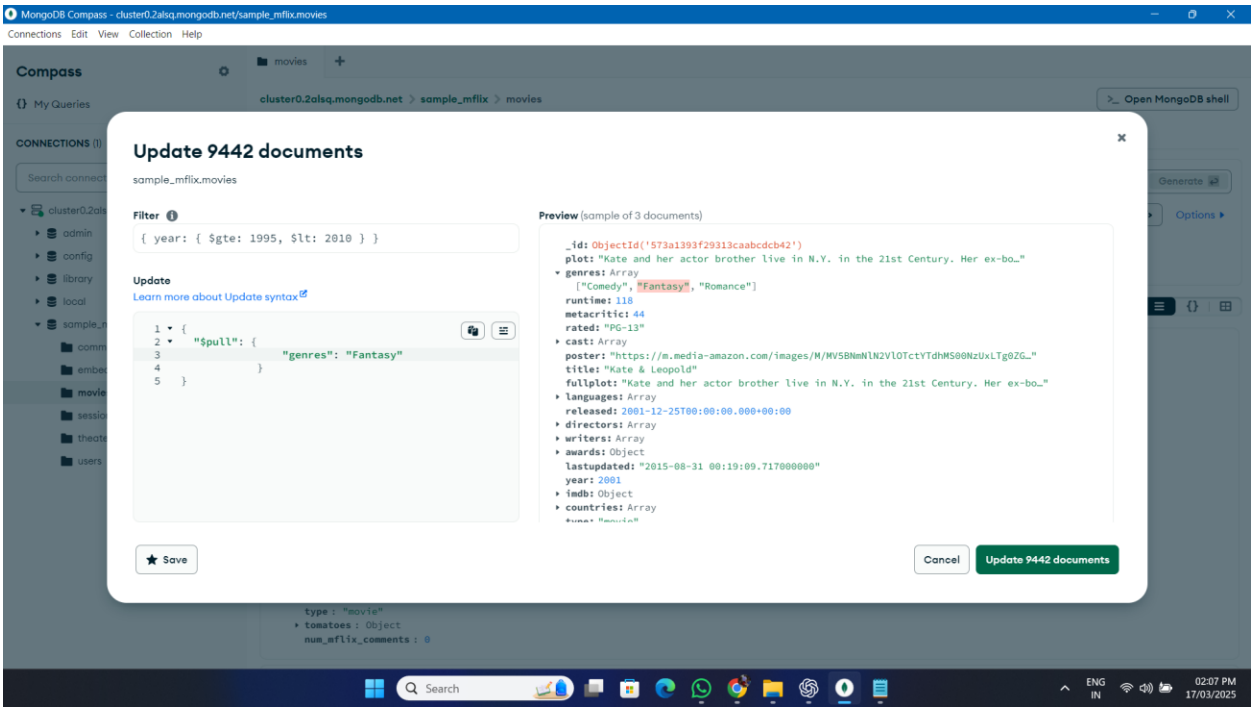
```
{
    "filter": {
        "runtime": {
            "$ne": null
        }
    },
    "update": {
        "$addToSet": {
            "awards.wins": 1
        }
    },
    "multi": true
}
```

5.3.1.



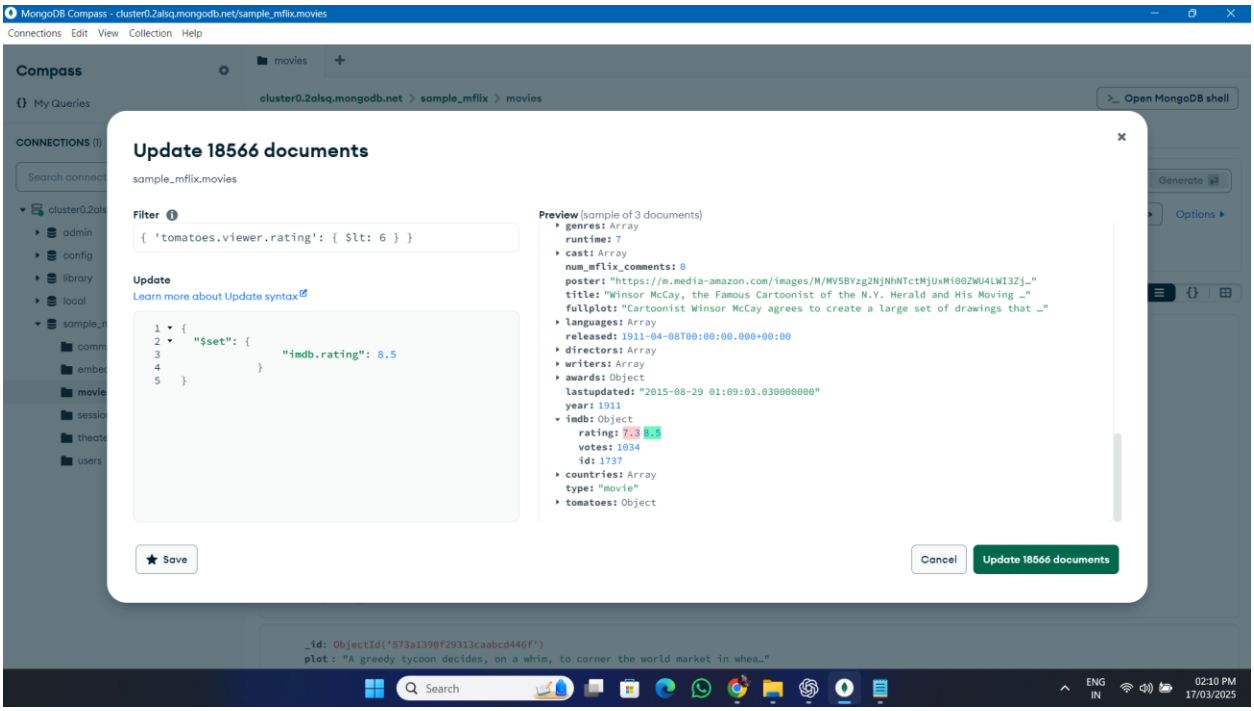
5.3.2.

```
{
      "filter": {
         "year": {
            "$gte": 1995,
            "$lt": 2010
         }
      },
      "update": {
         "$pull": {
            "genres": "Fantasy"
         }
      },
      "multi": true
   }
```



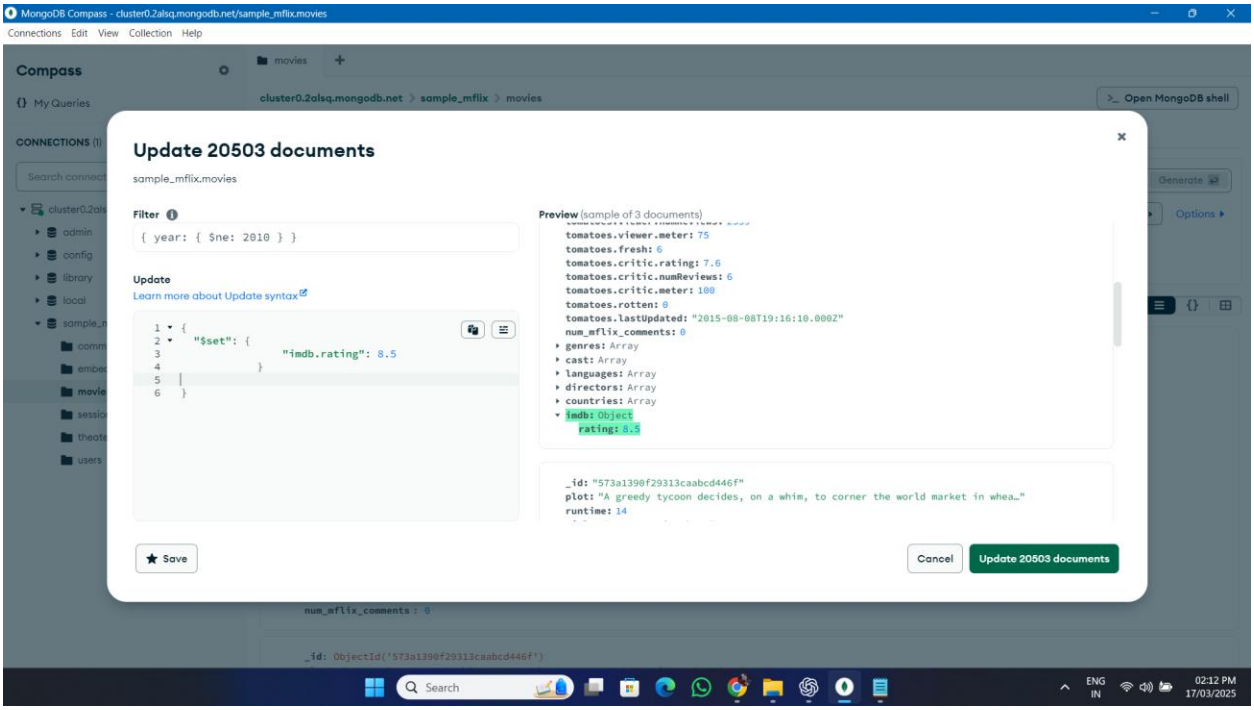
5.3.3.

```
{
      "filter": {
         "tomatoes.viewer.rating": {
            "$lt": 6.0
         }
      },
      "update": {
         "$set": {
            "imdb.rating": 8.5
         }
      },
      "multi": true
   }
```



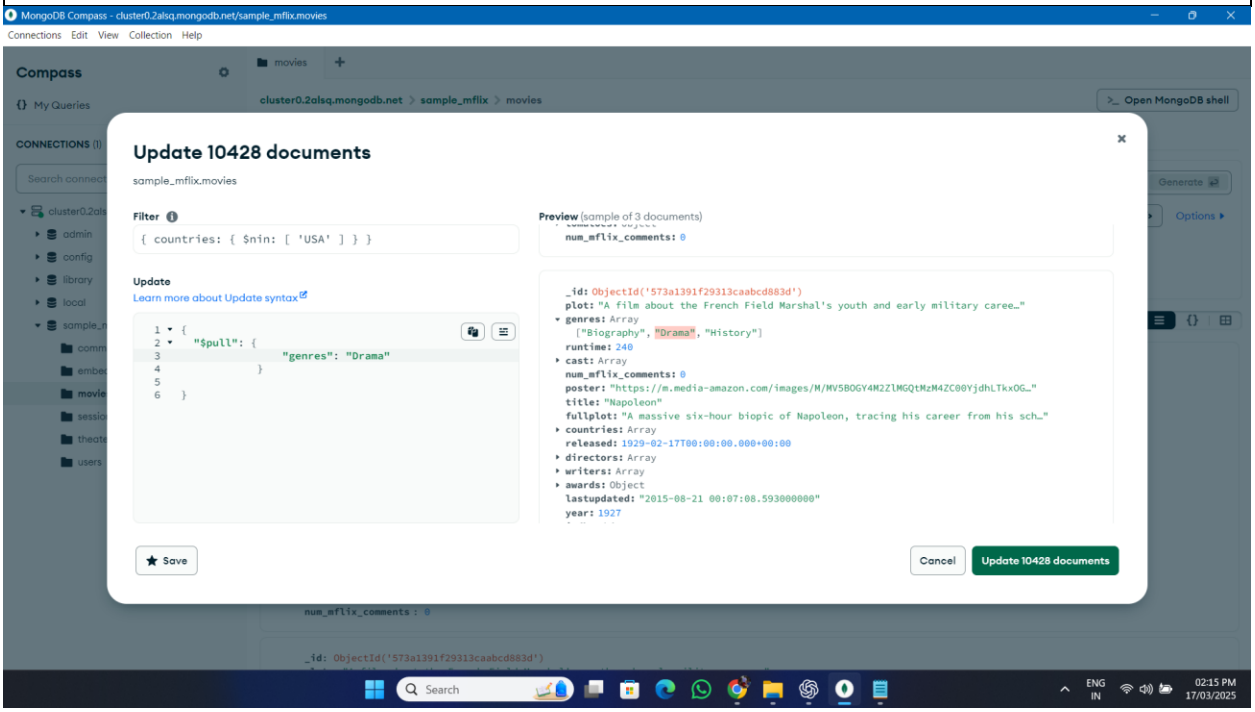
5.3.4.

```
{
      "filter": {
         "year": {
            "$ne": 2010
         }
      },
      "update": {
         "$set": {
            "imdb.rating": 8.5
         }
      },
      "multi": true
   }
```



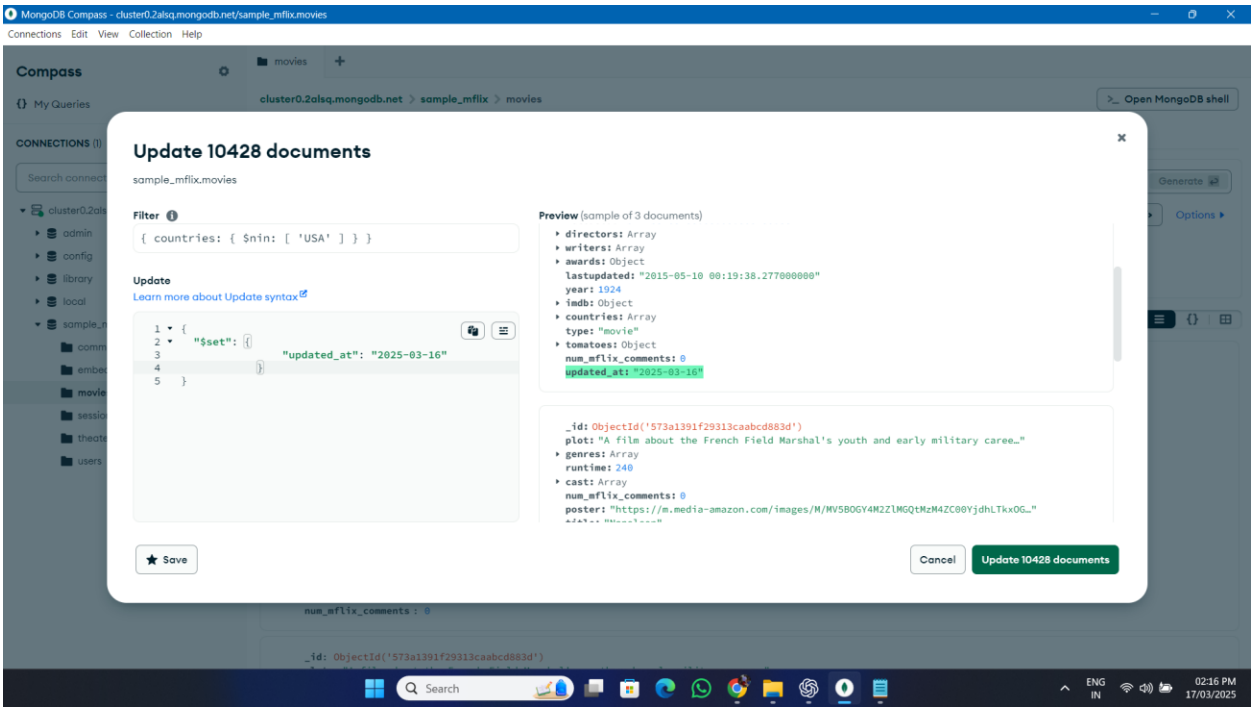
5.3.5.

```
{
      "filter": {
         "year": {
            "$ne": 2010
         }
      },
      "update": {
         "$set": {
            "imdb.rating": 8.5
         }
      },
      "multi": true
   }
```



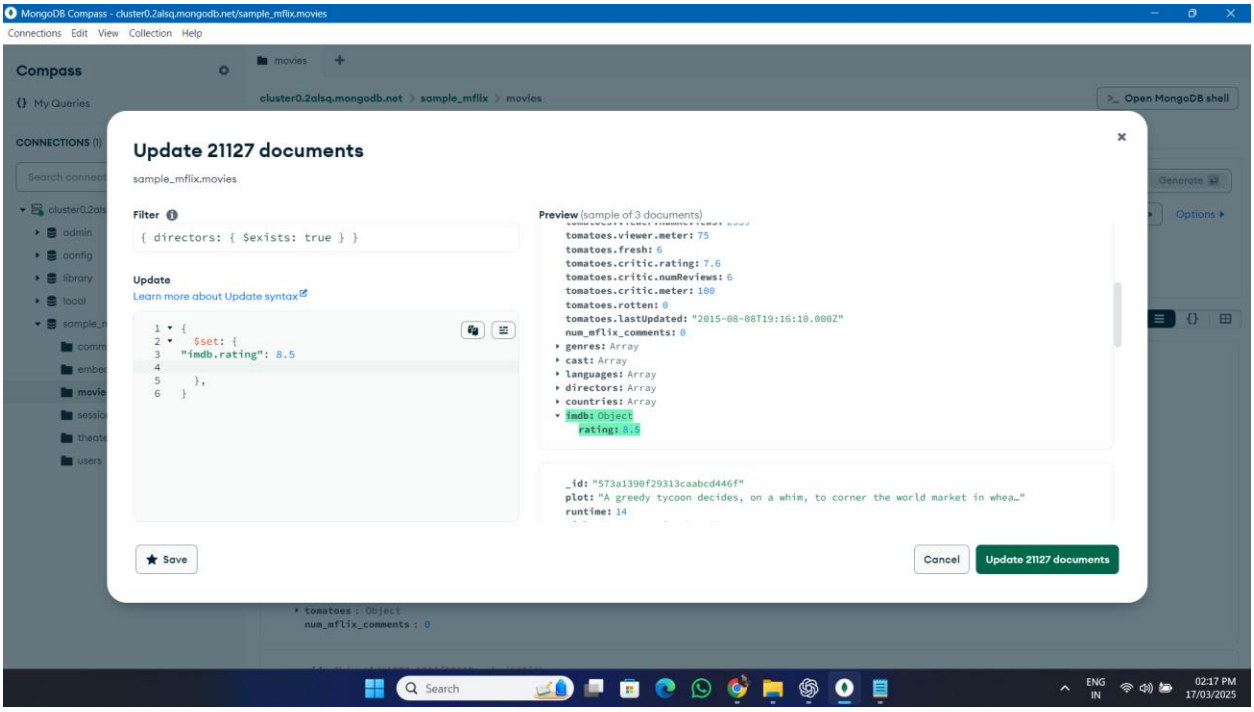
5.3.6.

```
{
      "filter": {
         "countries": {
            "$nin": [
               "USA"
            ]
         }
      },
      "update": {
         "$set": {
            "updated_at": "2025-03-16"
         }
      },
      "multi": true
   }
```



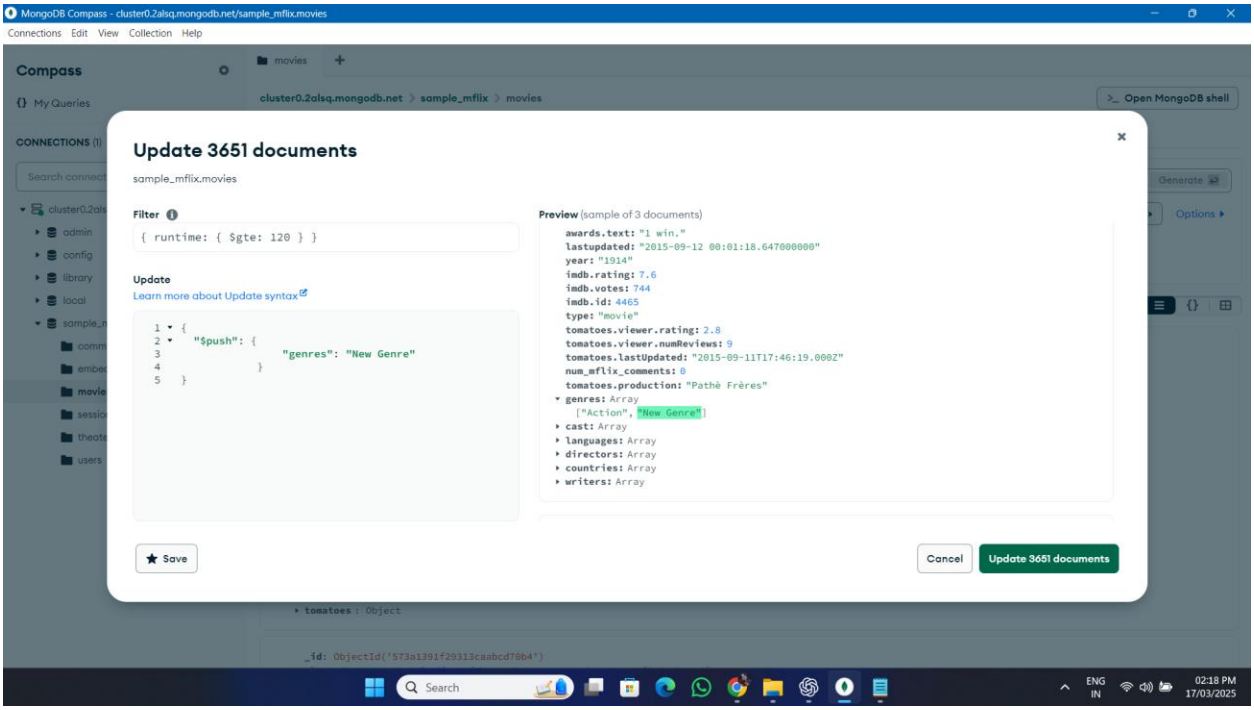
5.3.7.

```
{
      "filter": {
         "directors": {
            "$exists": true
         }
      },
      "update": {
         "$set": {
            "imdb.rating": 8.5
         }
      },
      "multi": true
   }
```



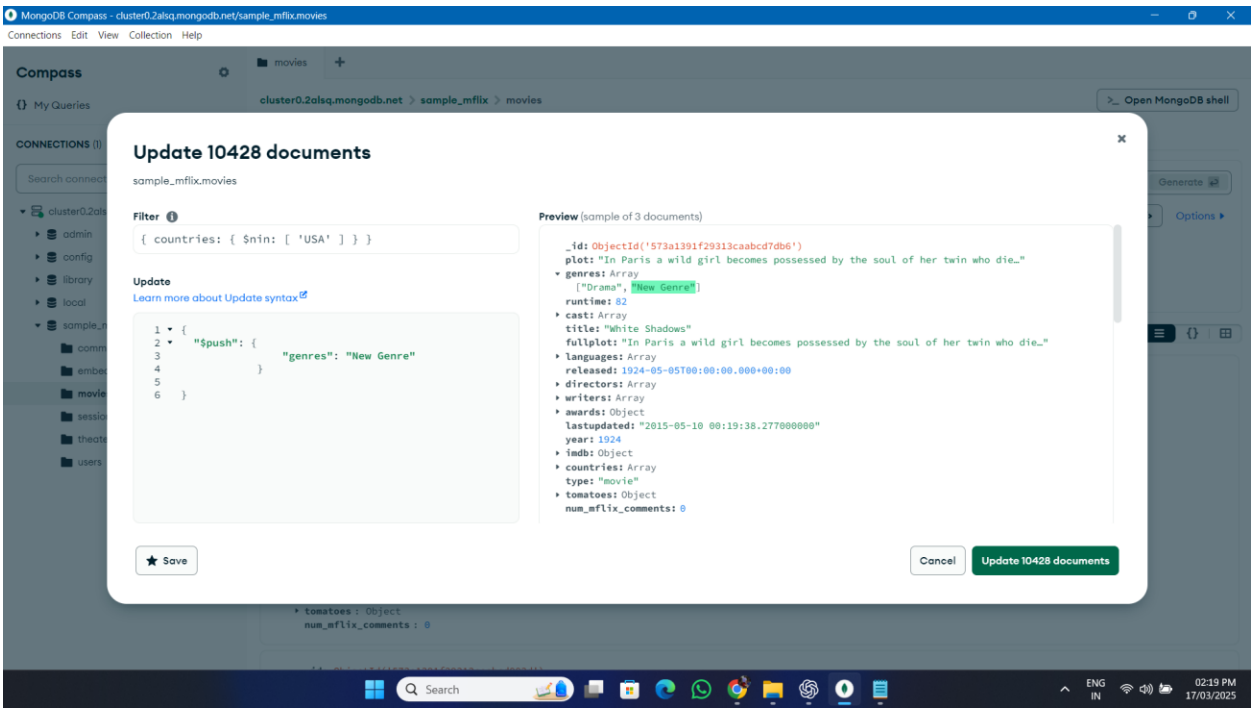
5.3.8.

```
{
      "filter": {
         "runtime": {
            "$gte": 120
         }
      },
      "update": {
         "$push": {
            "genres": "New Genre"
         }
      },
      "multi": true
   }
```



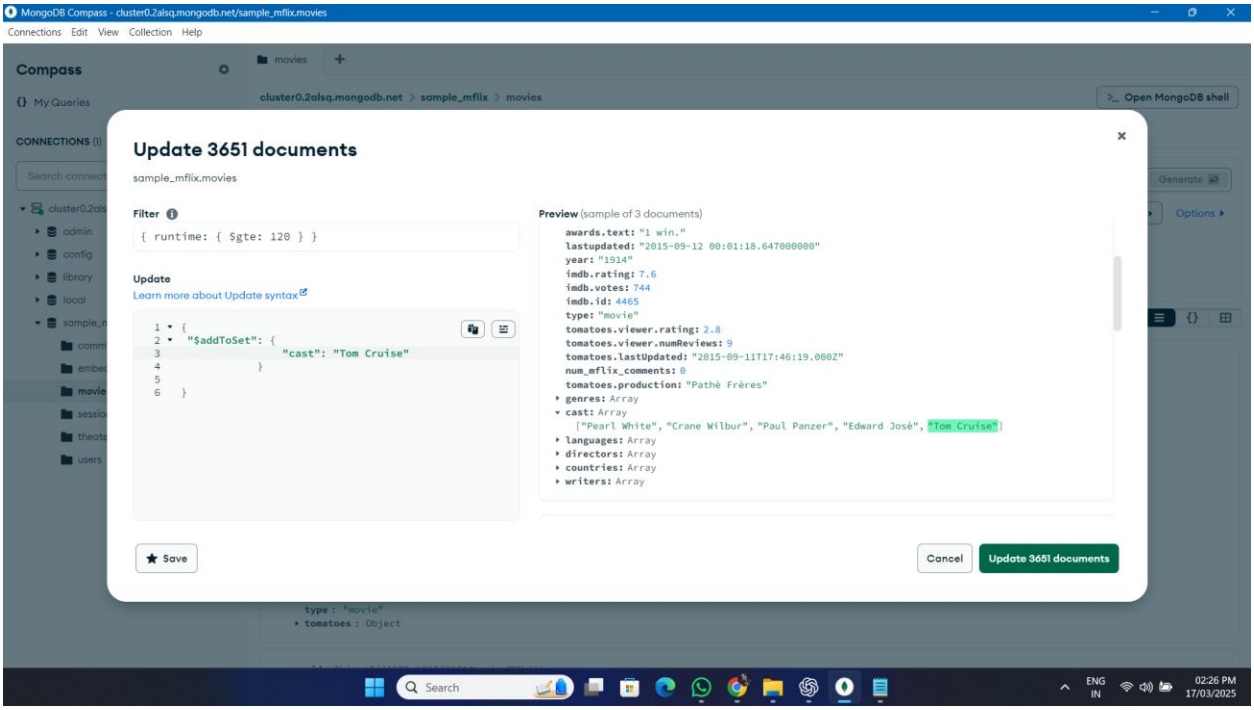
5.3.9.

```
{
      "filter": {
         "countries": {
            "$nin": [
               "USA"
            ]
         }
      },
      "update": {
         "$push": {
            "genres": "New Genre"
         }
      },
      "multi": true
   }
```



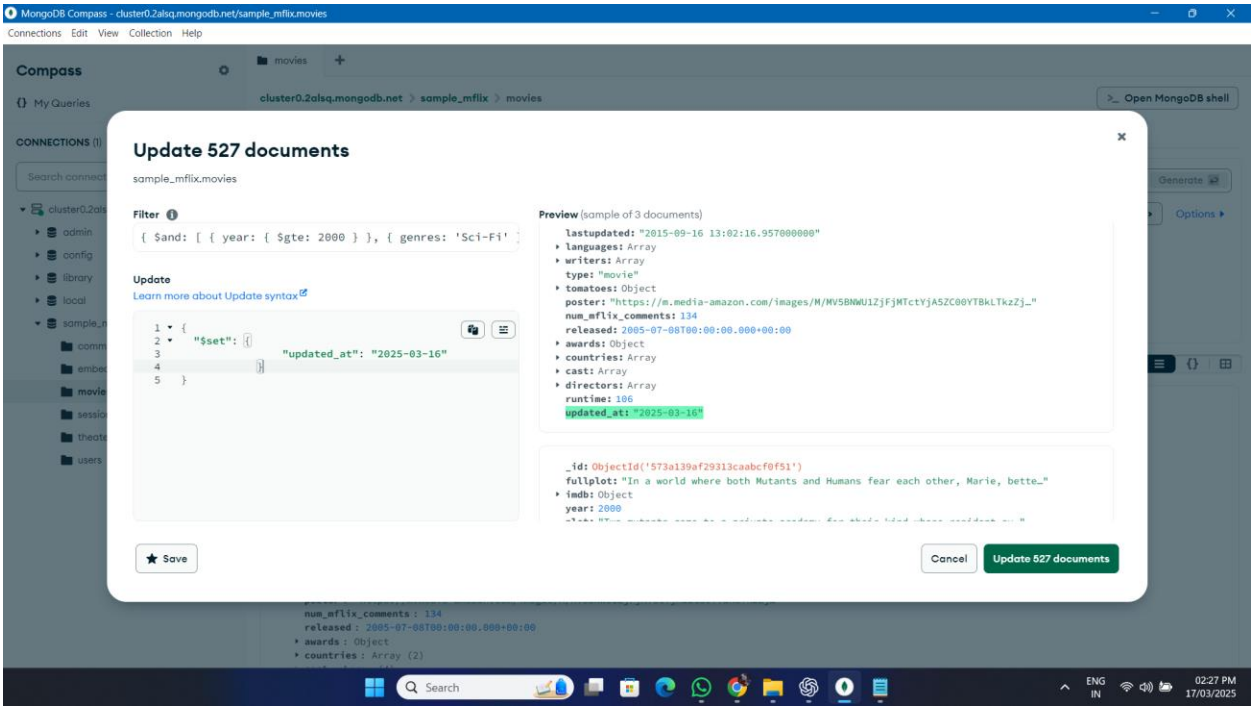
5.3.10.

```
{
      "filter": {
         "runtime": {
            "$gte": 120
         }
      },
      "update": {
         "$addToSet": {
            "cast": “Tom Cruise”
         }
      },
      "multi": true
   }
```



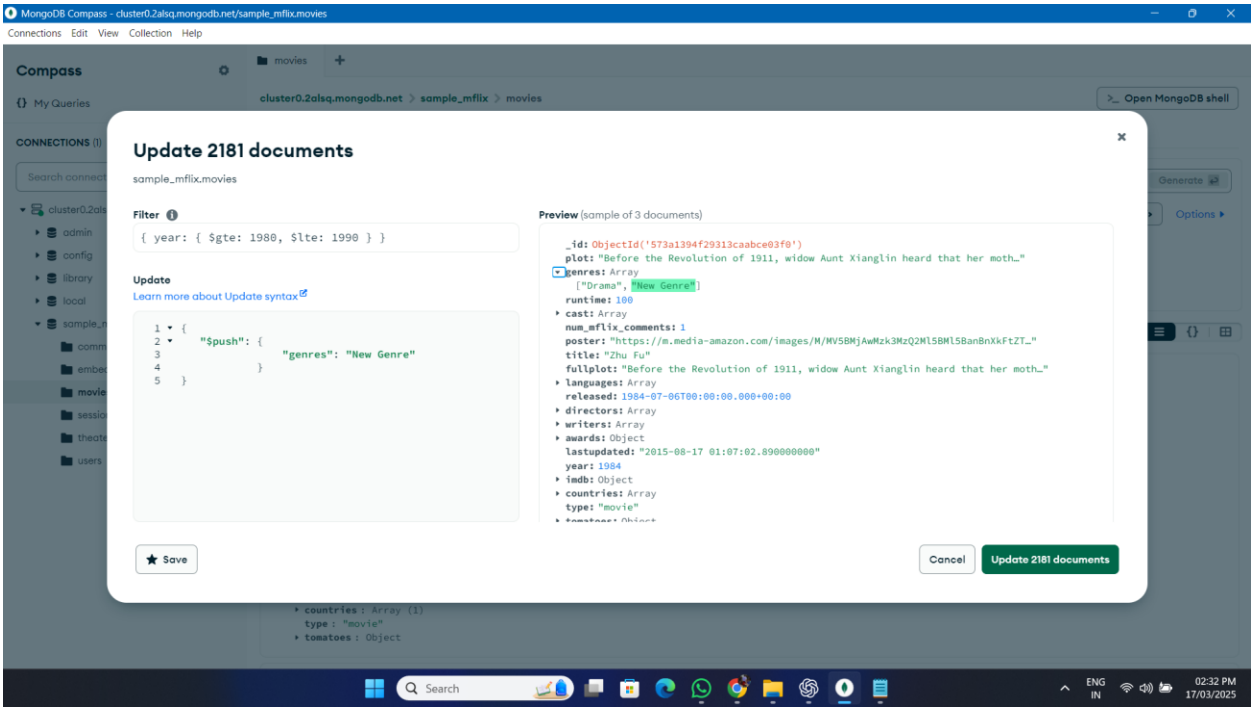
5.3.11.

```
{
      "filter": {
         "$and": [
            {
               "year": {
                  "$gte": 2000
               }
            },
            {
               "genres": "Sci-Fi"
            }
         ]
      },
      "update": {
         "$set": {
            "updated_at": "2025-03-16"
         }
      },
      "multi": true
   }
```



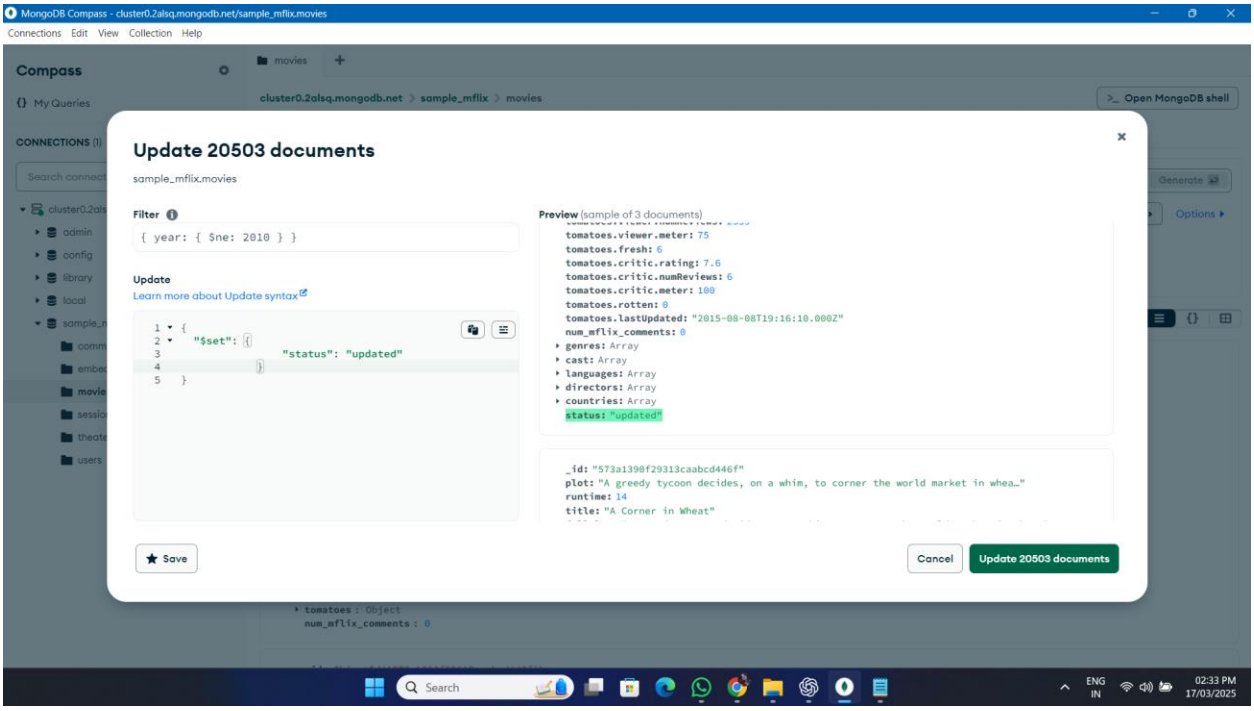
5.3.12.

```
{
      "filter": {
         "year": {
            "$gte": 1980,
            "$lte": 1990
         }
      },
      "update": {
         "$push": {
            "genres": "New Genre"
         }
      },
      "multi": true
   }
```



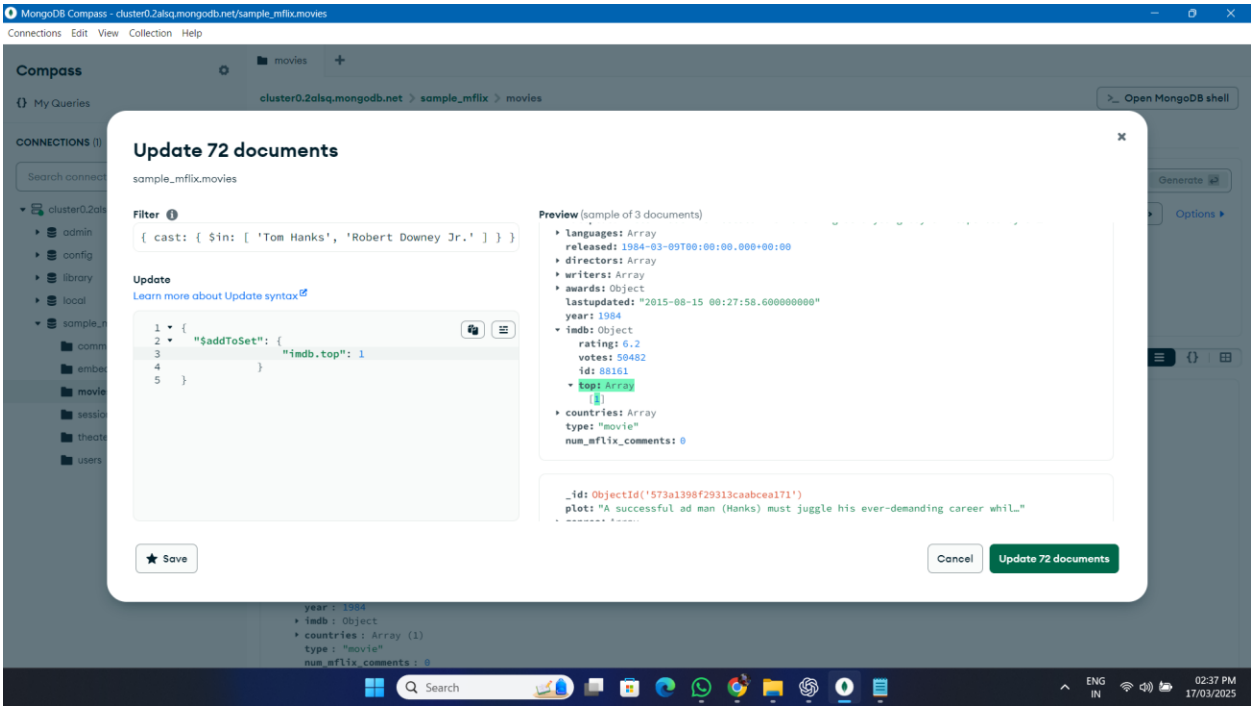
5.3.13.

```
{
      "filter": {
         "year": {
            "$ne": 2010
         }
      },
      "update": {
         "$set": {
            "status": "updated"
         }
      },
      "multi": true
   }
```



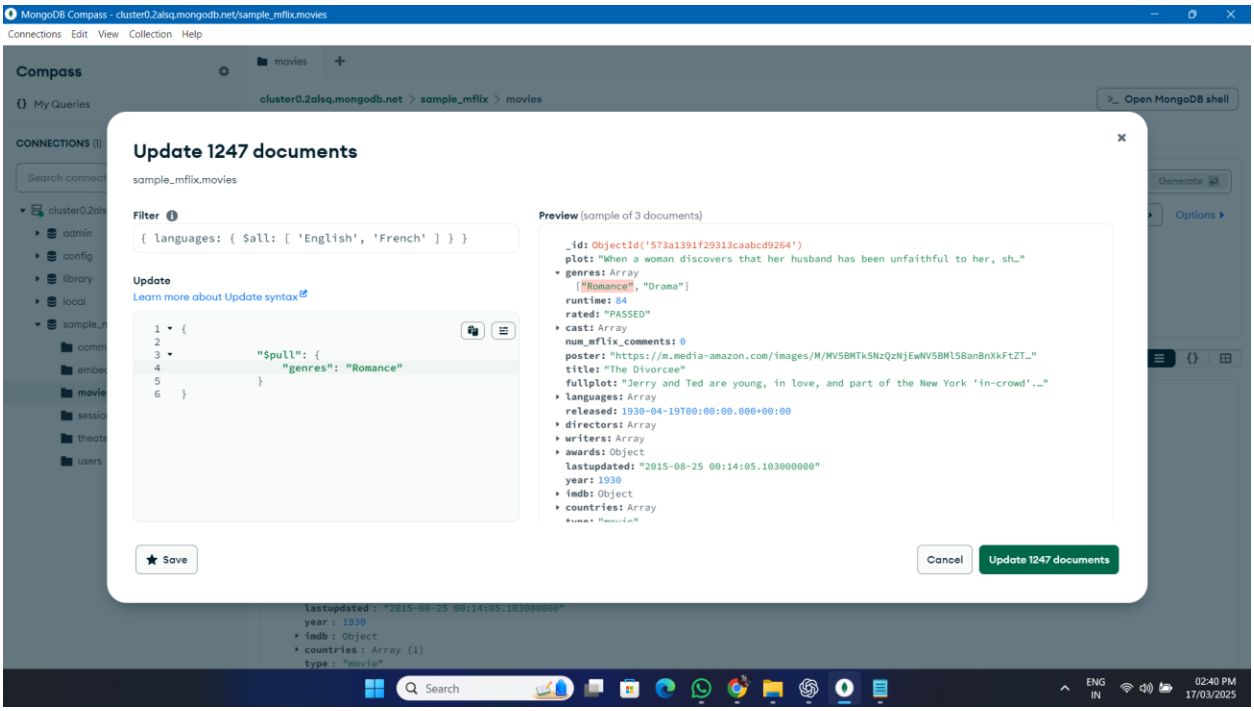
5.3.14.

```
{
      "filter": {
         "cast": {
            "$in": [
               "Tom Hanks",
               "Robert Downey Jr."
            ]
         }
      },
      "update": {
         "$addToSet": {
            "imdb.top": 1
         }
      },
      "multi": true
   }
```



5.3.15.

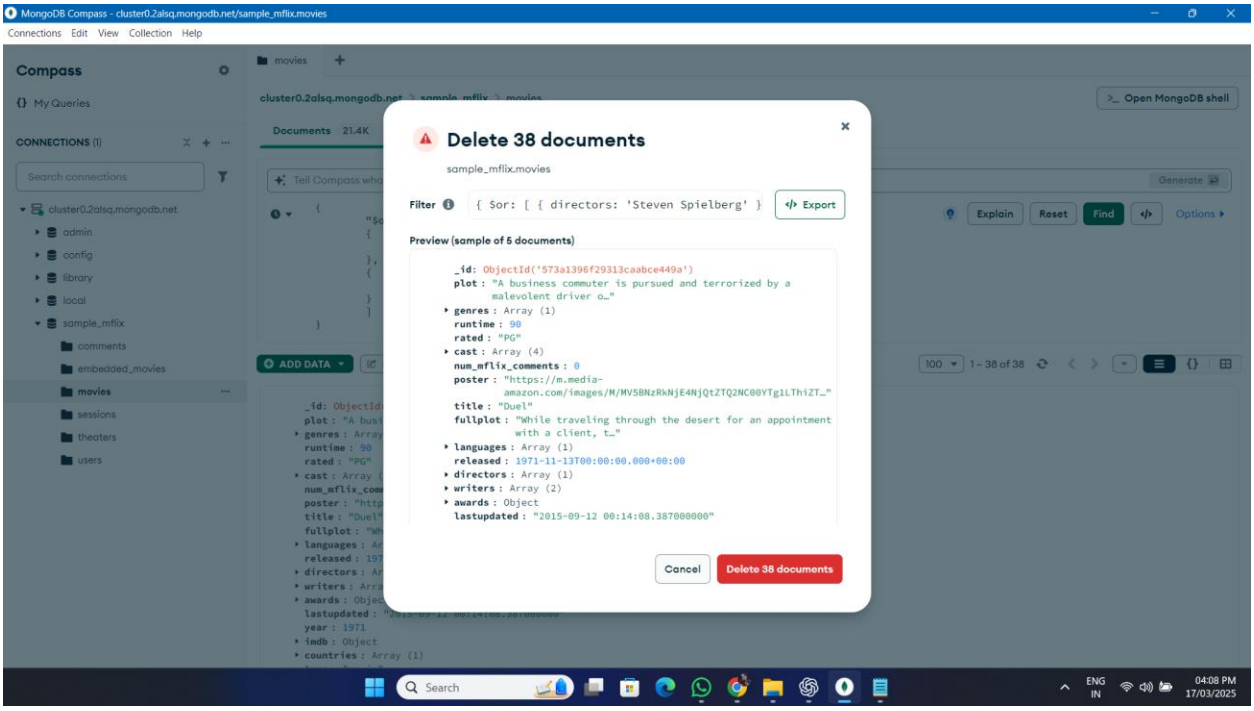
```
{
      "filter": {
         "languages": {
            "$all": [
               "English",
               "French"
            ]
         }
      },
      "update": {
         "$pull": {
            "genres": "Romance"
         }
      },
      "multi": true
   }
```



5.4. Delete

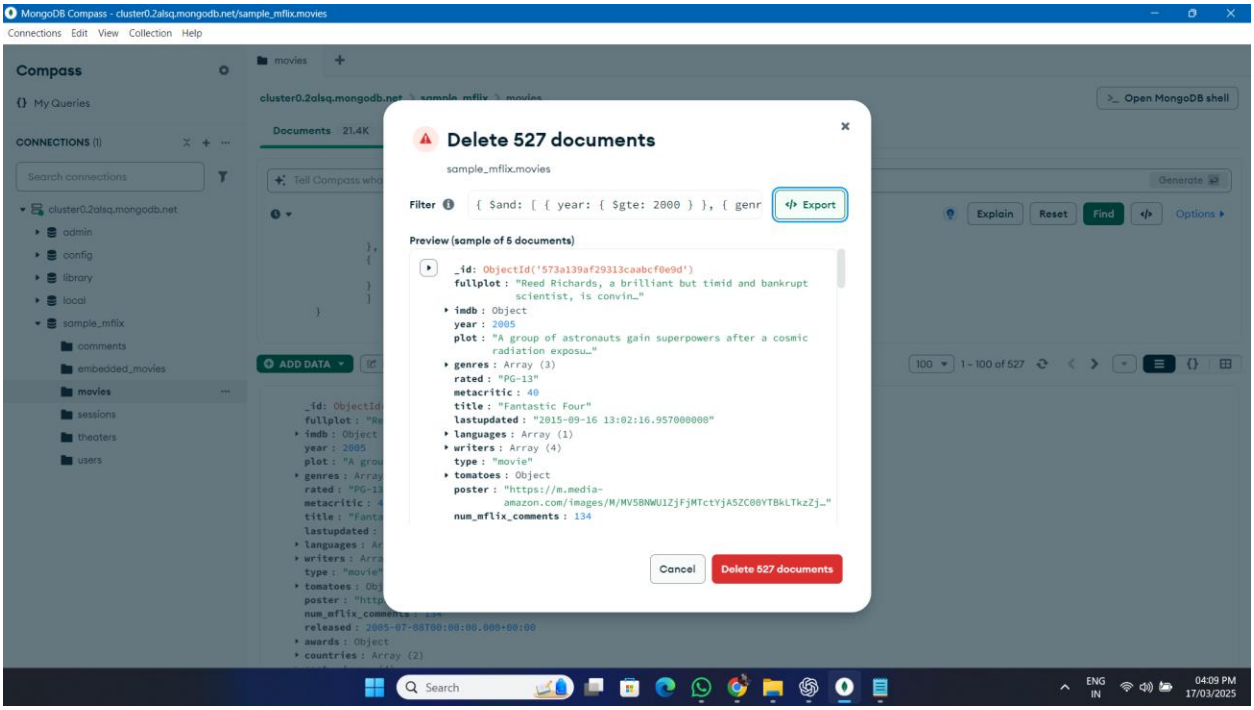
5.4.1.

```
{
      "$or": [
      {
            "directors": "Steven Spielberg"
      },
      {
            "directors": "Christopher Nolan"
      }
      ]
}
```



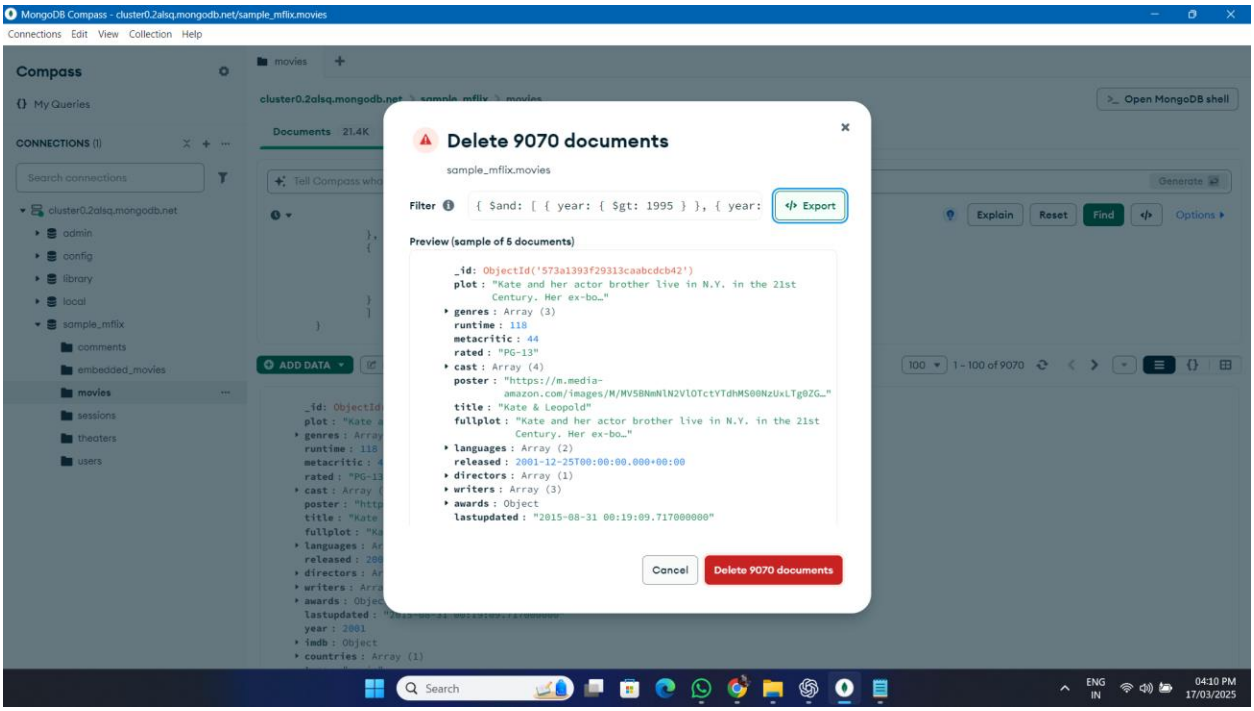
5.4.2.

```
{
      "$and": [
      {
            "year": {
               "$gte": 2000
         }
      },
      {
            "genres": "Sci-Fi"
      }
      ]
}
```



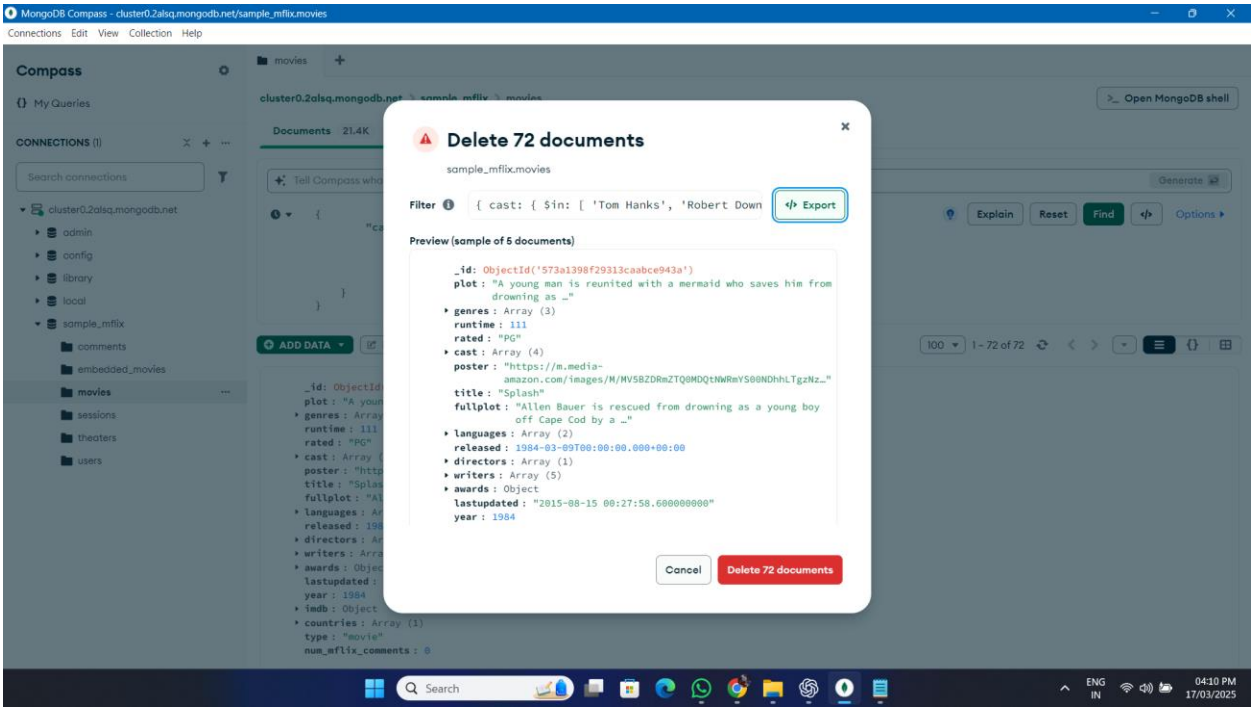
5.4.3.

```
{
      "$and": [
      {
            "year": {
               "$gt": 1995
         }
      },
      {
            "year": {
               "$lt": 2010
         }
      }
      ]
}
```



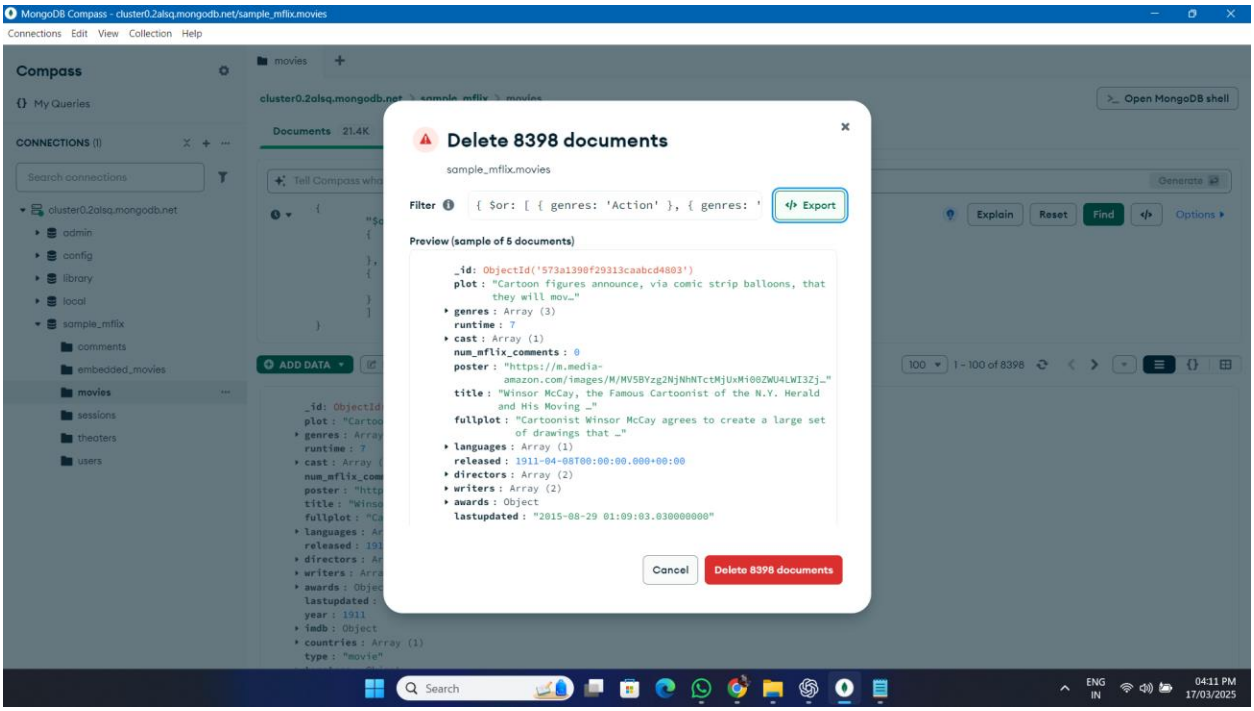
5.4.4.

```
{
      "cast": {
         "$in": [
            "Tom Hanks",
            "Robert Downey Jr."
         ]
   }
}
```



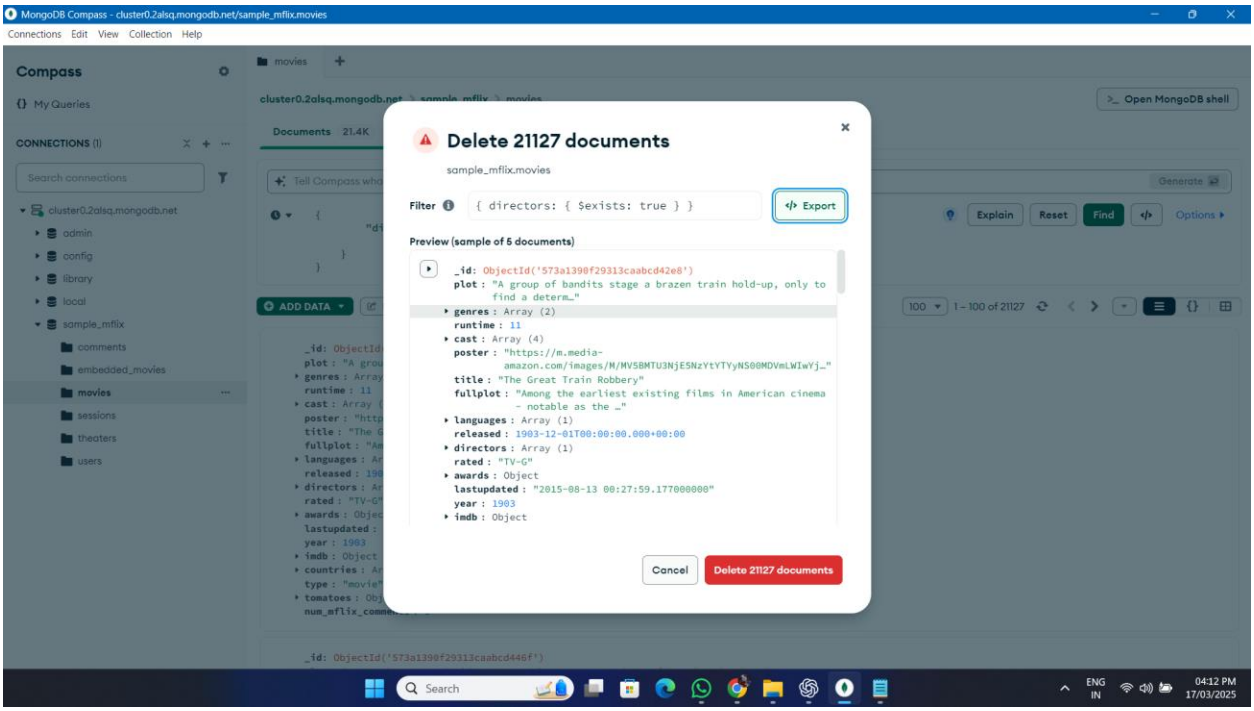
5.4.5.

```
{
      "$or": [
      {
            "genres": "Action"
      },
      {
            "genres": "Comedy"
      }
      ]
}
```



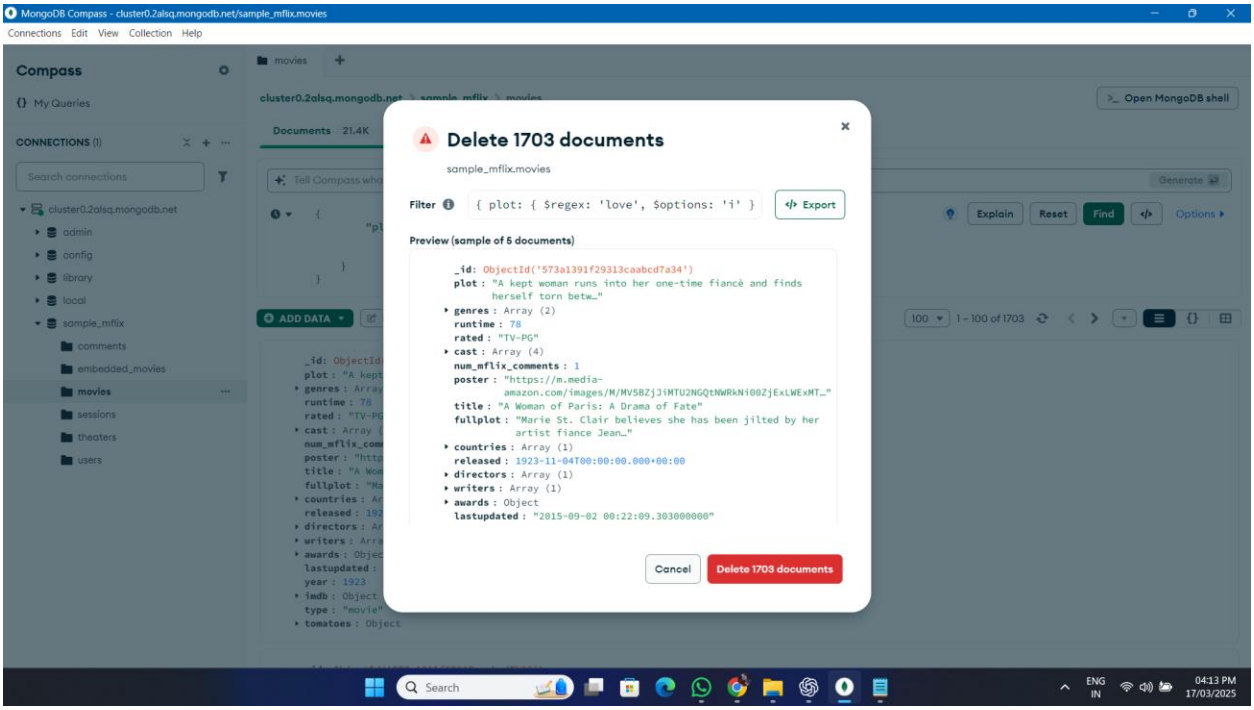
5.4.6.

```
{
      "directors": {
         "$exists": true
   }
}
```



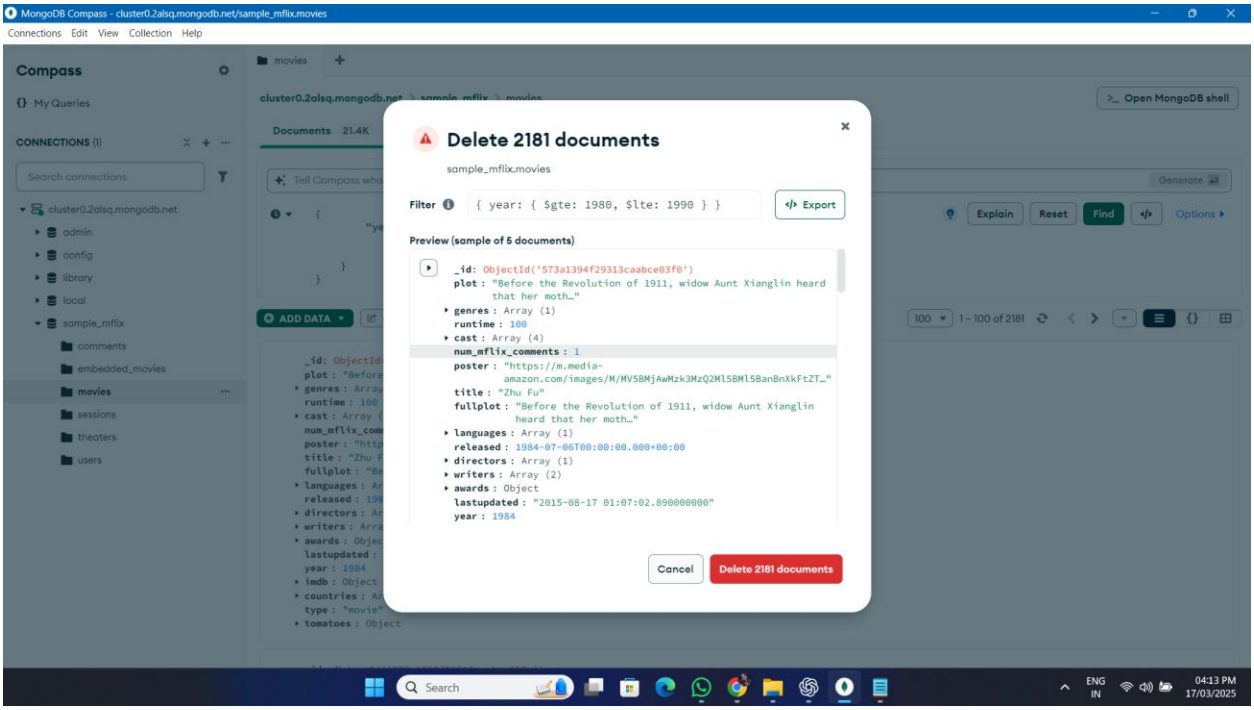
5.4.7.

```
{
      "plot": {
         "$regex": "love",
         "$options": "i"
   }
}
```



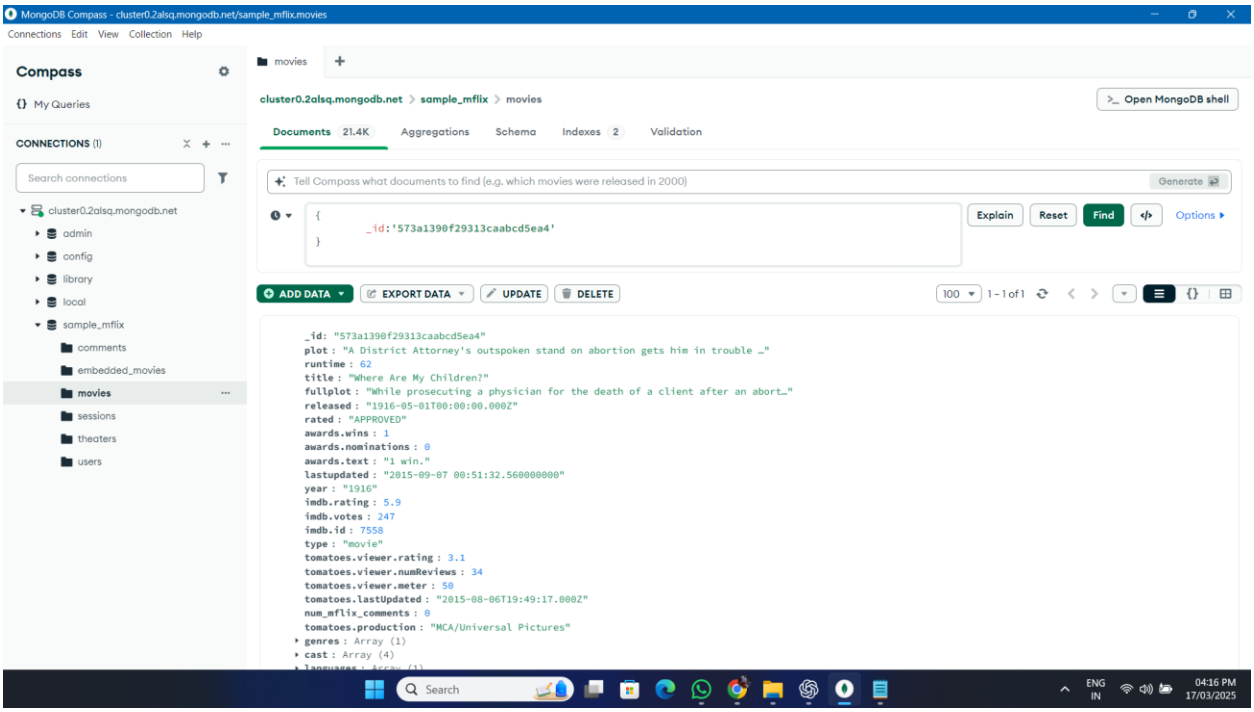
5.4.8.

```
{
      "year": {
         "$gte": 1980,
         "$lte": 1990
   }
}
```



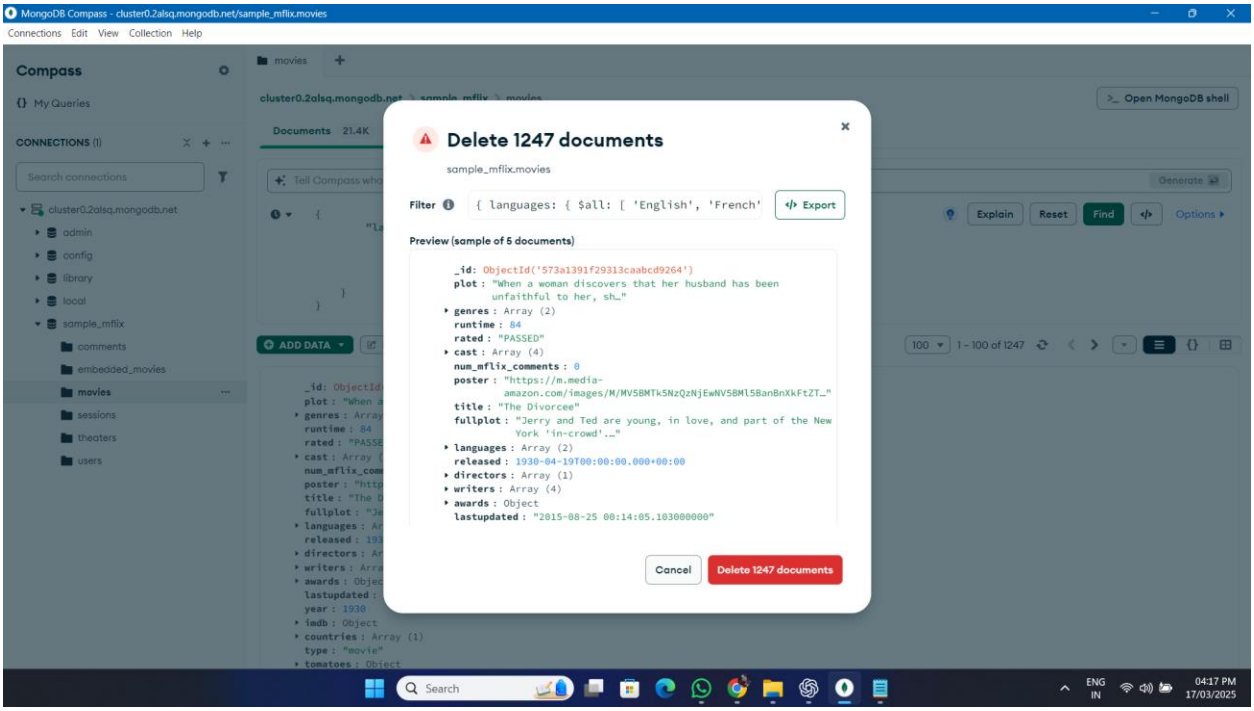
5.4.9.

```
{
        _id:’573a1390f29313caabcd5ea4’
}
```



5.4.10.

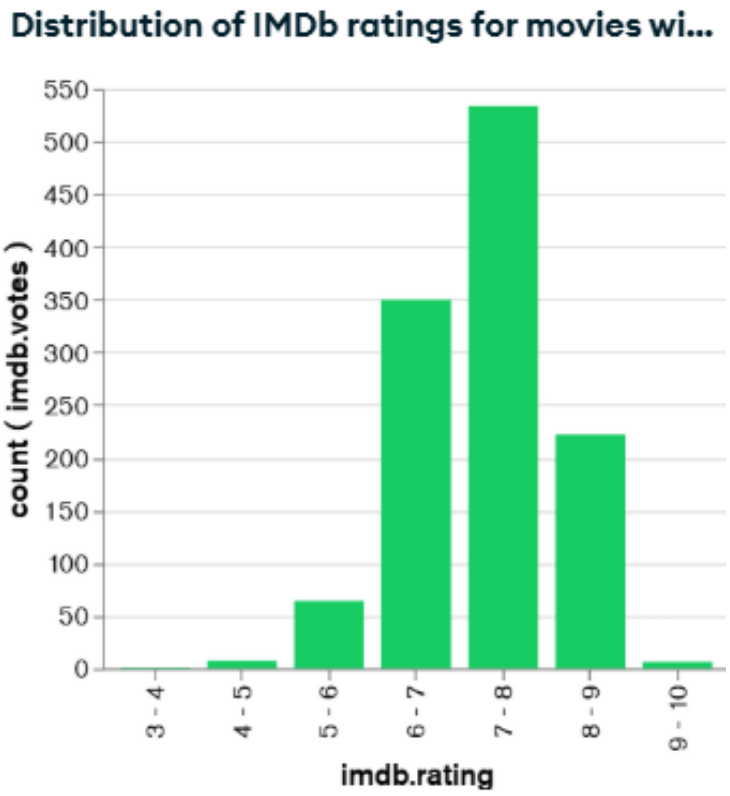
```
{
      "languages": {
         "$all": [
            "English",
            "French"
         ]
   }
}
```



5.5. Key Observations and Insights

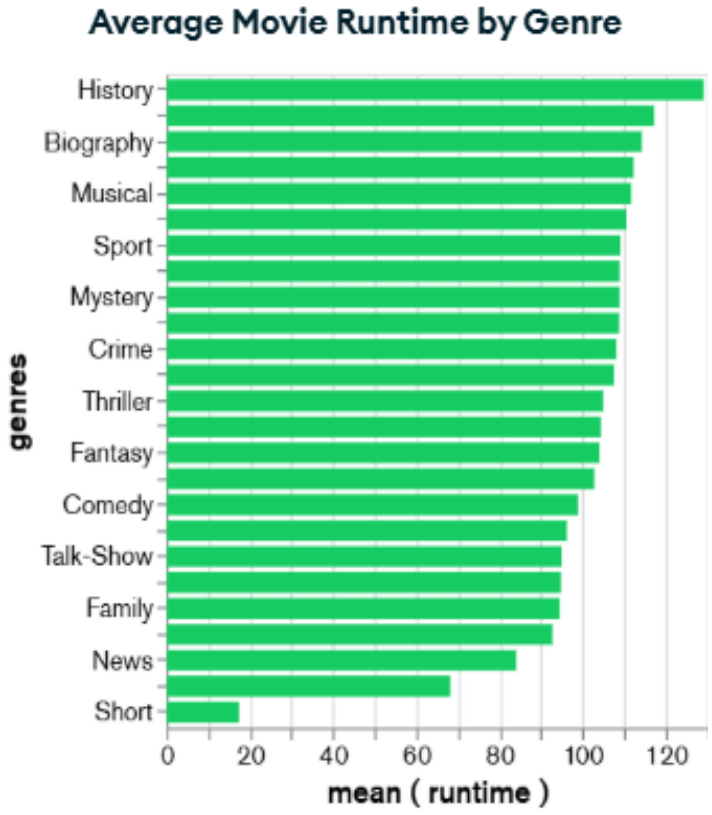
5.5.1. Distribution of IMDb Ratings for Movies with More Than 100,000 Votes

- Most movies fall within the **7.0–8.0 rating range**, indicating that audiences generally rate popular movies positively.
- A significant drop is observed for movies rated **below 6.0**, showing that fewer widely popular movies receive lower ratings.
- Very few movies have ratings below **4.0 or above 9.0**, suggesting that extreme ratings (very low or very high) are rare.



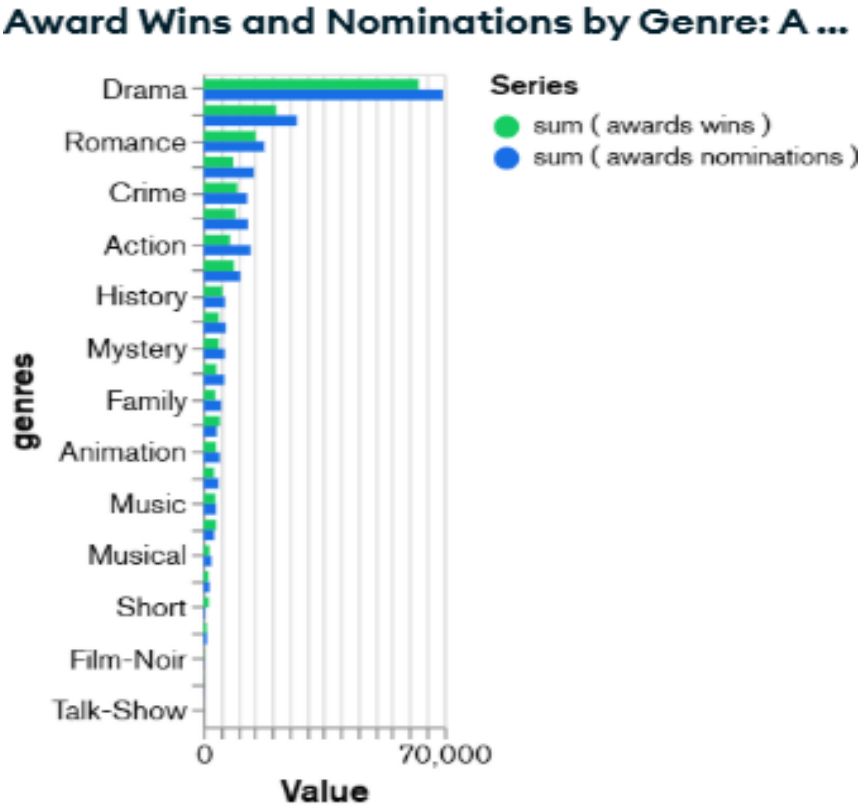
5.5.2. Average Movie Runtime by Genre

- **Historical movies have the longest runtime**, averaging over 120 minutes, which aligns with the genre's need for extensive storytelling.
- **Short films have the lowest runtime**, as expected, reinforcing the distinction in duration between different formats.
- Biographical, musical, and sports genres also tend to have **longer runtimes**, possibly due to the need for detailed character development.



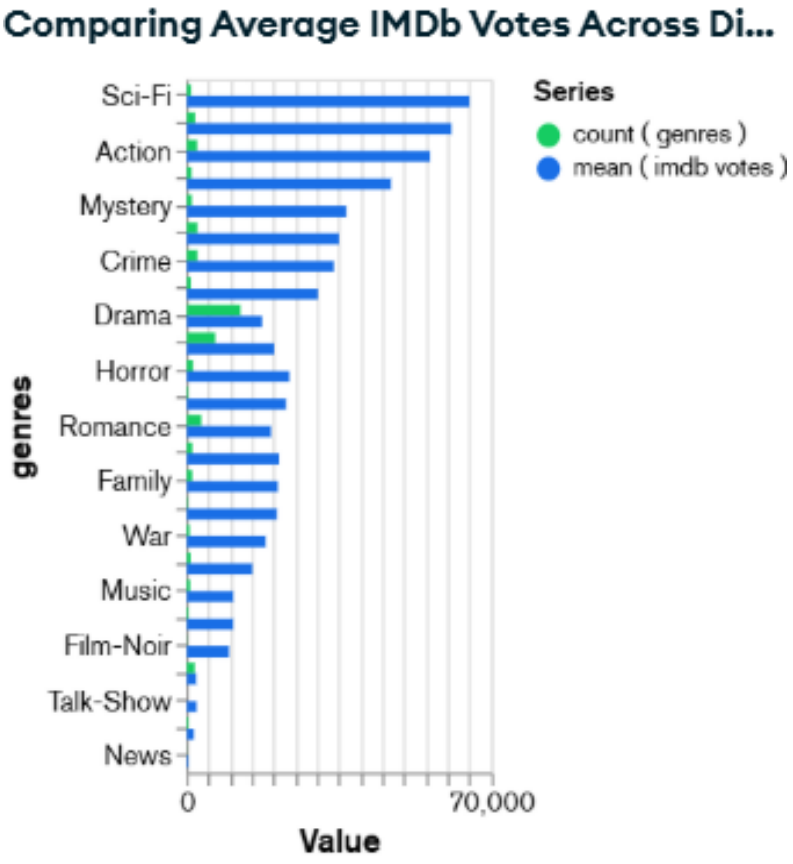
5.5.3. Award Wins and Nominations by Genre: A Comparative Analysis

- **Drama dominates** in award wins and nominations, suggesting that this genre is highly favored by award committees.
- **Romance, Crime, and Action** follow in recognition, although their numbers are significantly lower than Drama.
- **Genres like Talk-Show and Film-Noir** receive very few awards, indicating they may not be mainstream choices for major award bodies.



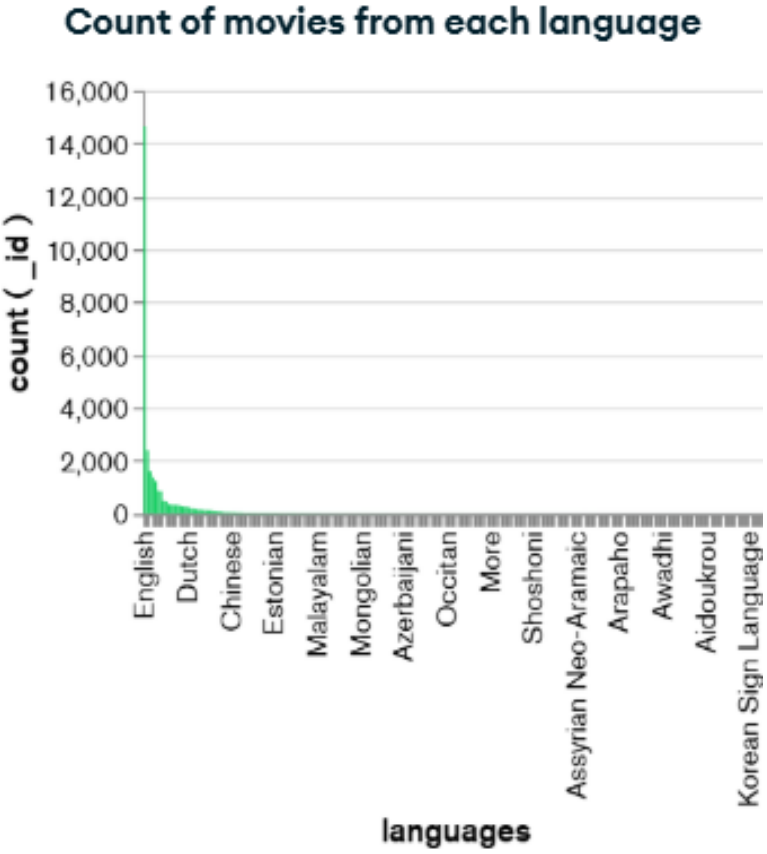
5.5.4. Comparing Average IMDb Votes Across Different Film Genres

- **Sci-Fi and Action** genres receive the **highest average IMDb votes**, suggesting these genres attract a larger audience and fan engagement.
- **Drama, Horror, and Mystery** also perform well, showing that audience interest spans across both intense and emotional storytelling.
- **Genres like Talk-Show and News** receive significantly fewer votes, indicating a niche audience rather than widespread appeal.



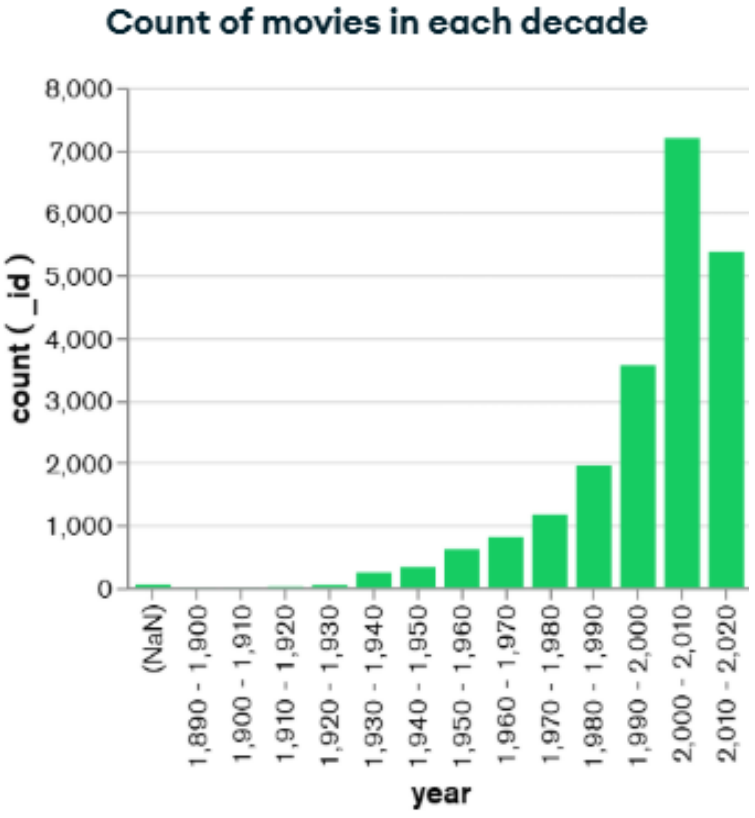
5.5.5. Count of Movies from Each Language

- **English dominates** the dataset, with a significantly higher number of movies compared to any other language.
- The distribution follows a **long-tail pattern**, where a few languages have a high number of films, but most languages have very few.
- **Languages like Dutch, Chinese, and Estonian** have a noticeable presence, but they are far behind English.



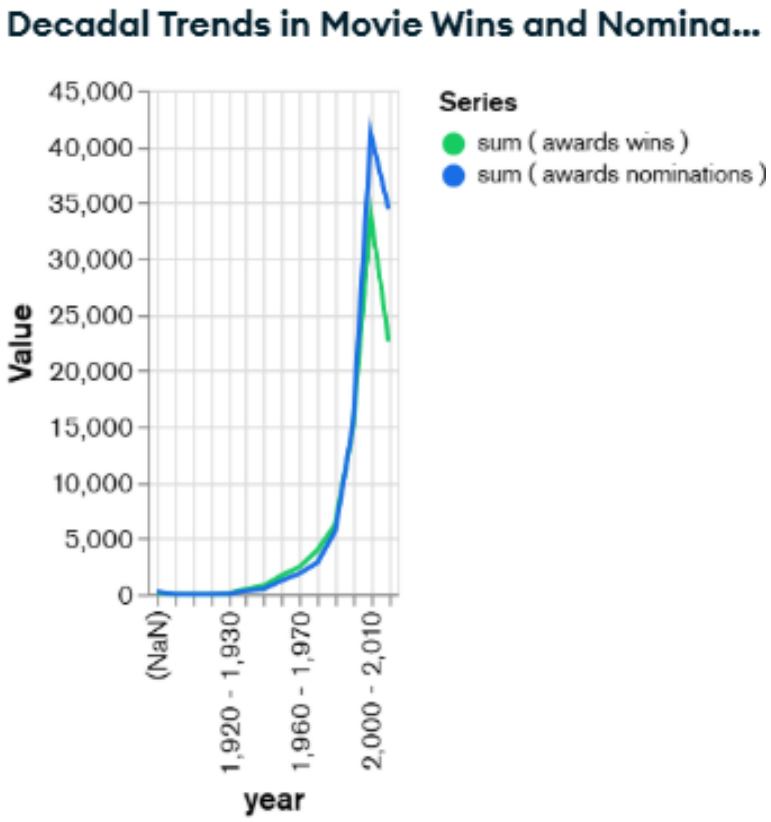
5.5.6. Count of Movies in Each Decade

- The number of movies has **increased drastically over time**, with the highest count in the **2000s and 2010s**.
- Early decades (1890–1950) have **very few movies**, indicating limited production capacity and technology.
- The rise in movie production **post-1980** aligns with technological advancements and the growth of the entertainment industry.



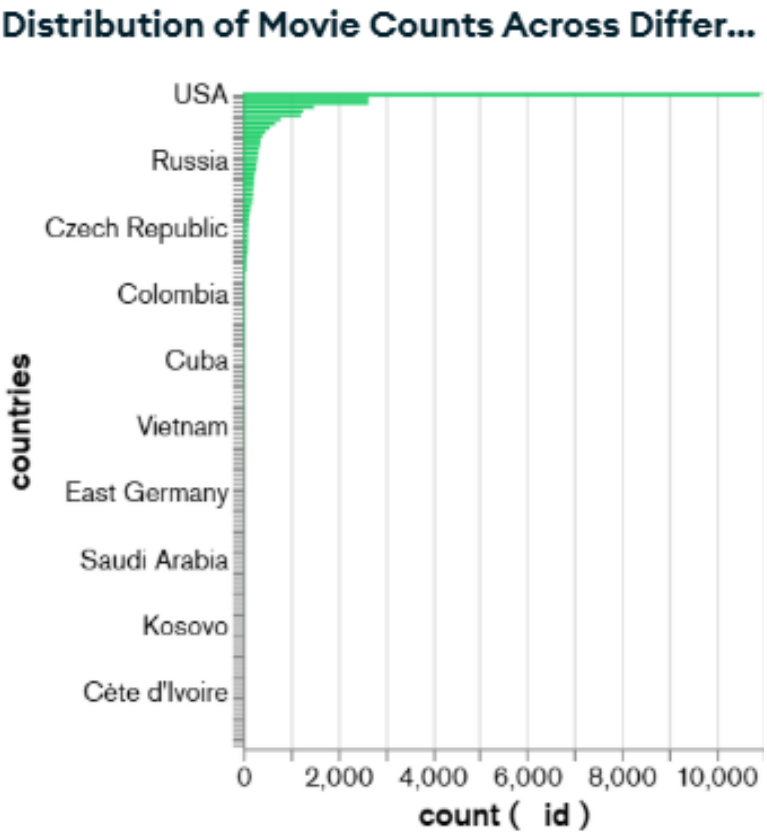
5.5.7. Decadal Trends in Movie Wins and Nominations

- Award wins and nominations have **increased significantly since the 1970s**, peaking in the 2000s.
- This may indicate a **rise in film festivals, streaming content, and award events**.



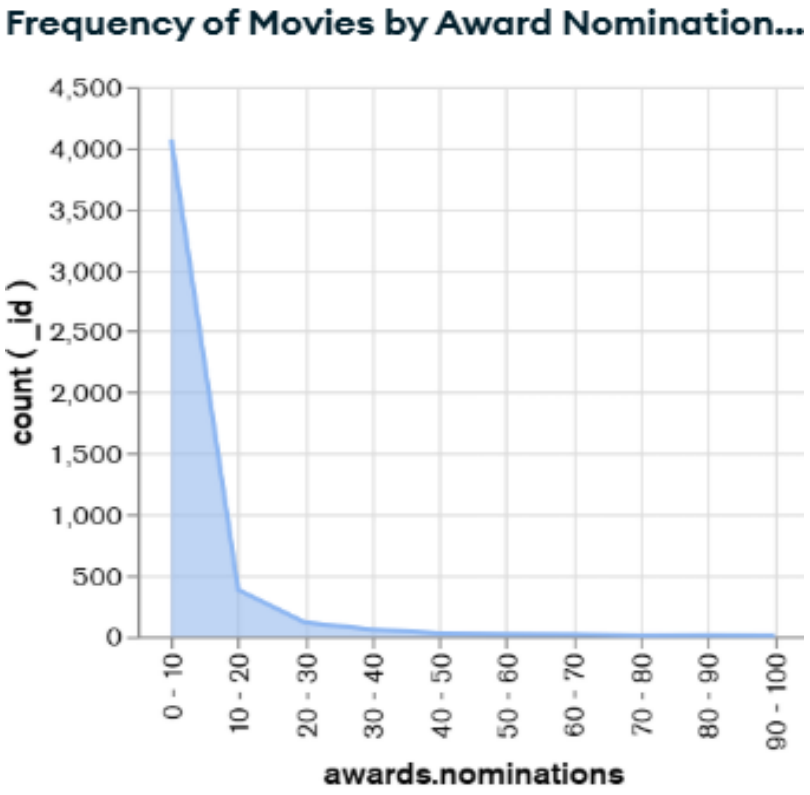
5.5.8. Distribution of Movie Counts Across Different Countries

- **The USA** has an overwhelmingly high number of movies compared to other countries.
- **Russia** follows, but no other country comes close, highlighting Hollywood's **dominant influence** in film production.



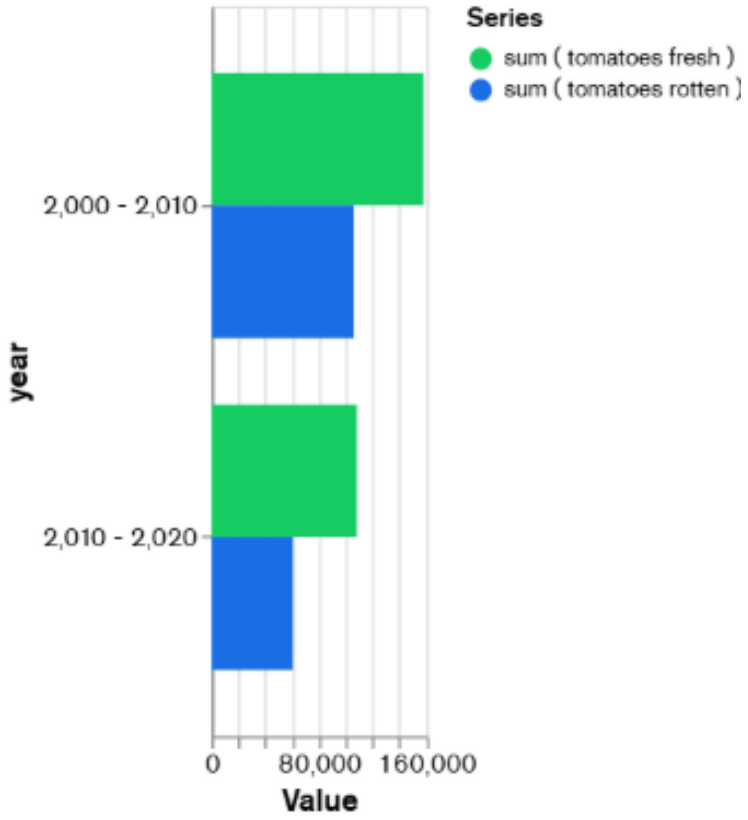
5.5.9. Frequency of Movies by Award Nominations Count

- Most movies receive **fewer than 10 nominations**, while very few movies get high nominations (above 50).
- This suggests that **highly nominated films are rare** and prestigious.



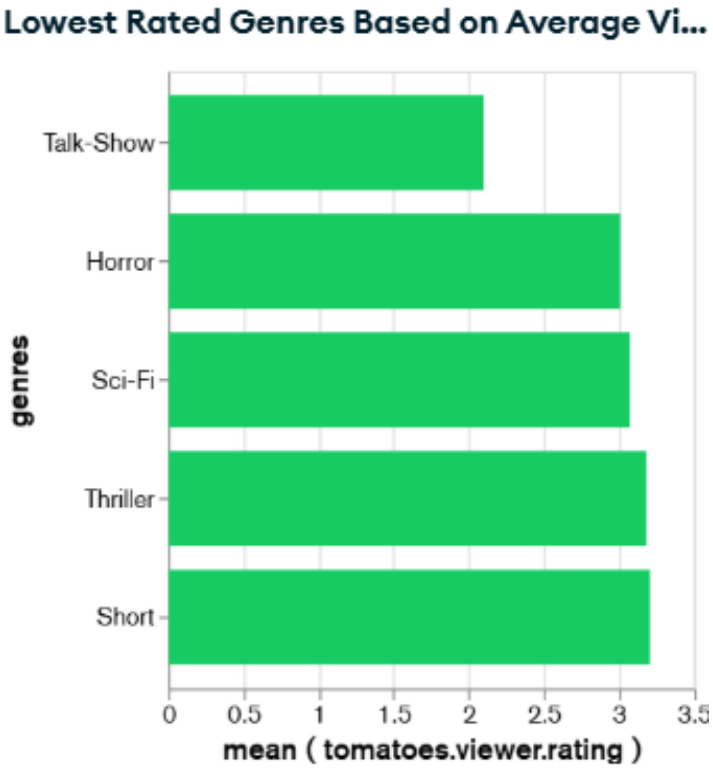
5.5.10. Grouped Bar Chart of Rotten Tomatoes Ratings (Fresh vs. Rotten)

- More movies received **"Fresh" ratings** than "Rotten," suggesting a general tendency for movies to be rated positively.
- However, there is still a **substantial number** of negatively rated movies.



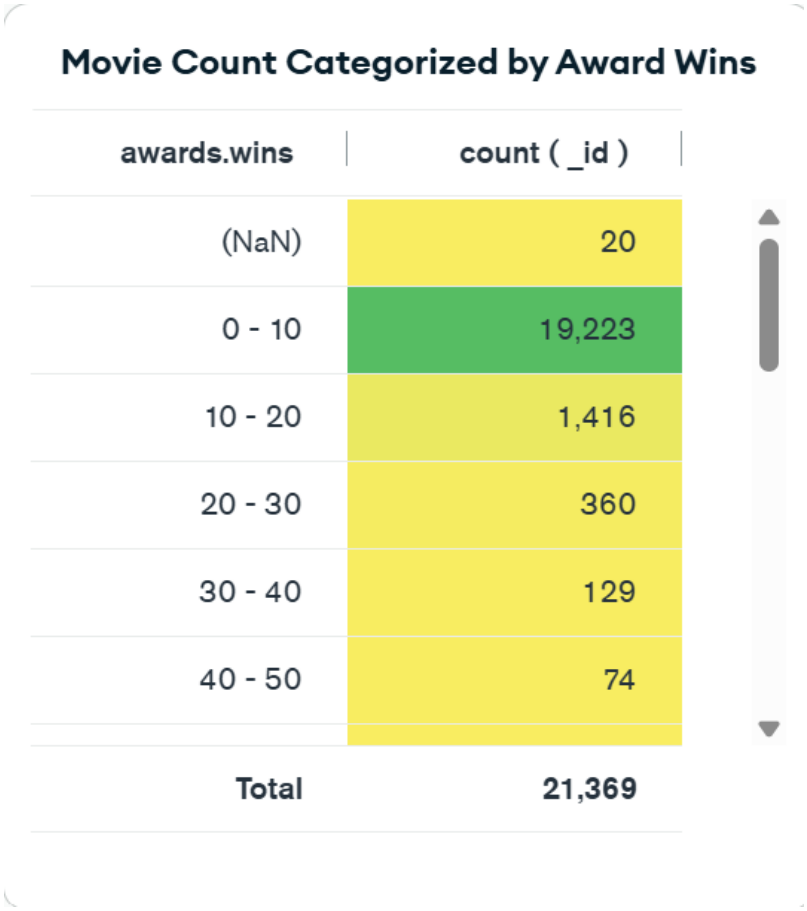
5.5.11. Lowest Rated Genres Based on Average Viewer Ratings

- **Talk-Show, Horror, and Sci-Fi** have the lowest average viewer ratings.
- **Thriller and Short** films also rank low, indicating that audiences might find them less appealing.



5.5.12. Movie Count Categorized by Award Wins

- The majority of movies (over **19,000**) have won between **0-10 awards**.
- Only **a few hundred** movies have won more than **20 awards**, making highly awarded films extremely rare.

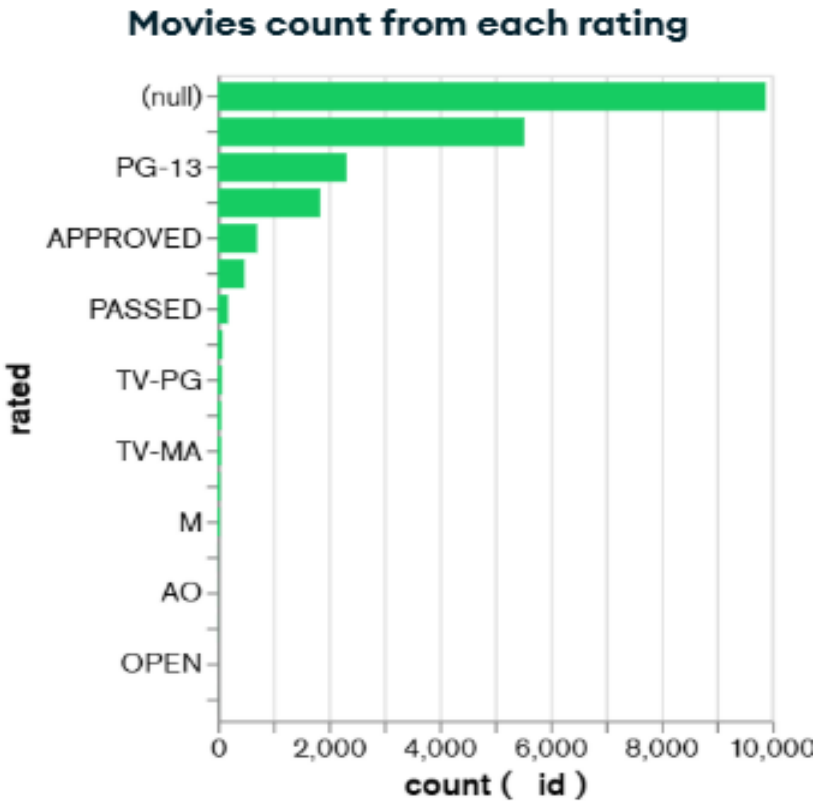


Movie Count Categorized by Award Wins

awards.wins	count (_id)
(NaN)	20
0 - 10	19,223
10 - 20	1,416
20 - 30	360
30 - 40	129
40 - 50	74
Total	**21,369**

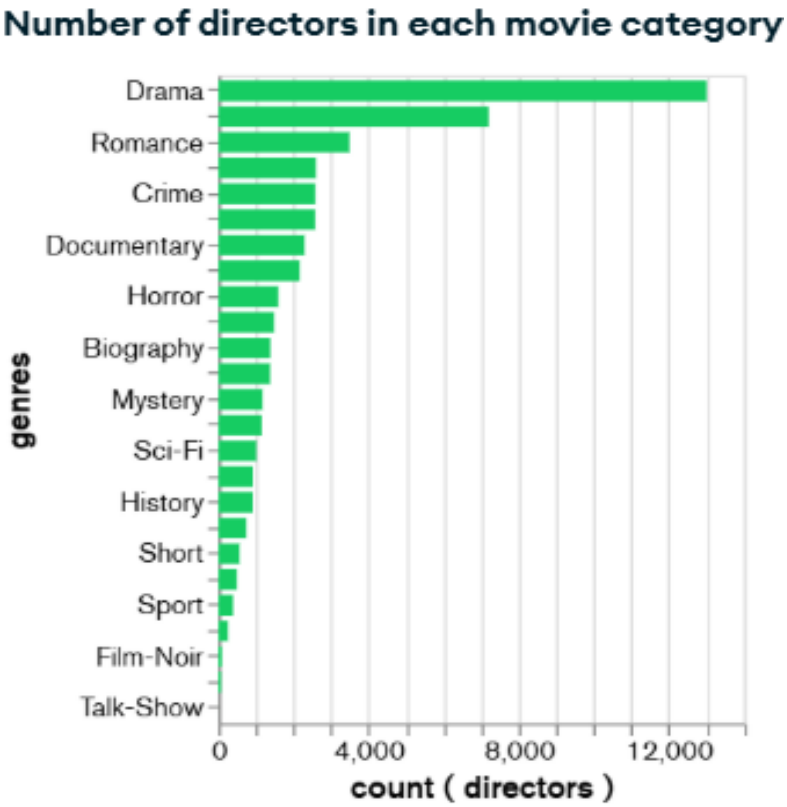
5.5.13. Movies Count from Each Rating

- A significant number of movies **lack a specified rating** (null), indicating missing or unclassified rating data.
- The most common explicit rating is **PG-13**, suggesting that a large portion of movies cater to a wide audience.
- Older rating systems like **APPROVED** and **PASSED** have significantly fewer movies, indicating that most films in the dataset are from the modern rating system.
- Ratings such as **TV-MA** and **AO** have very low counts, suggesting that explicit or adult content films are less frequent in this dataset.



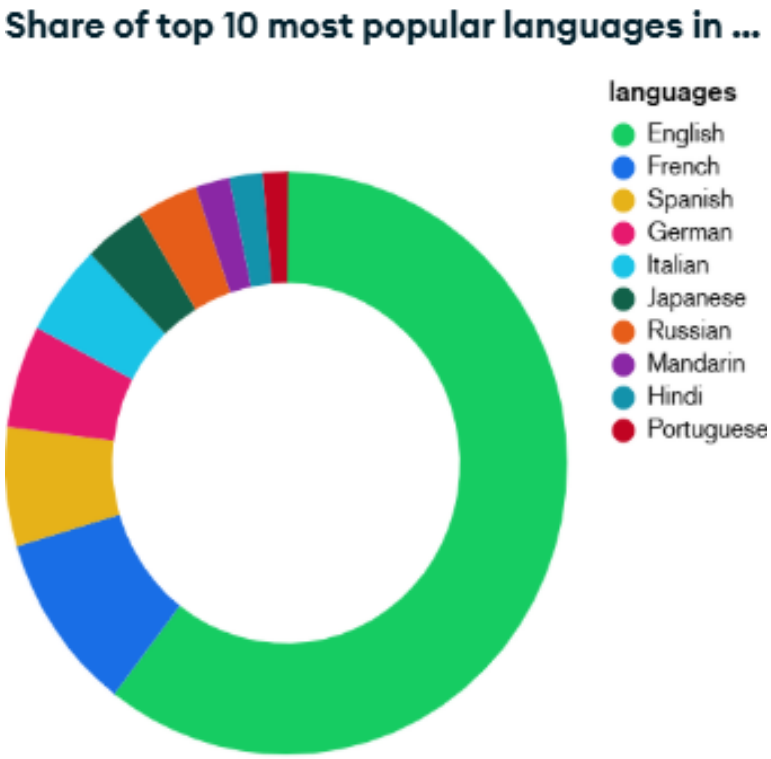
5.5.14. Number of Directors in Each Movie Category

- **Drama** has the highest number of directors, indicating its popularity and the variety of storytelling in this genre.
- **Romance and Crime** also have significant director counts, showing their strong presence in the film industry.
- **Documentary** has a high director count, suggesting a strong representation of real-world narratives in filmmaking.
- Niche genres like **Film-Noir and Talk-Show** have the lowest number of directors, which aligns with their relatively smaller market share.



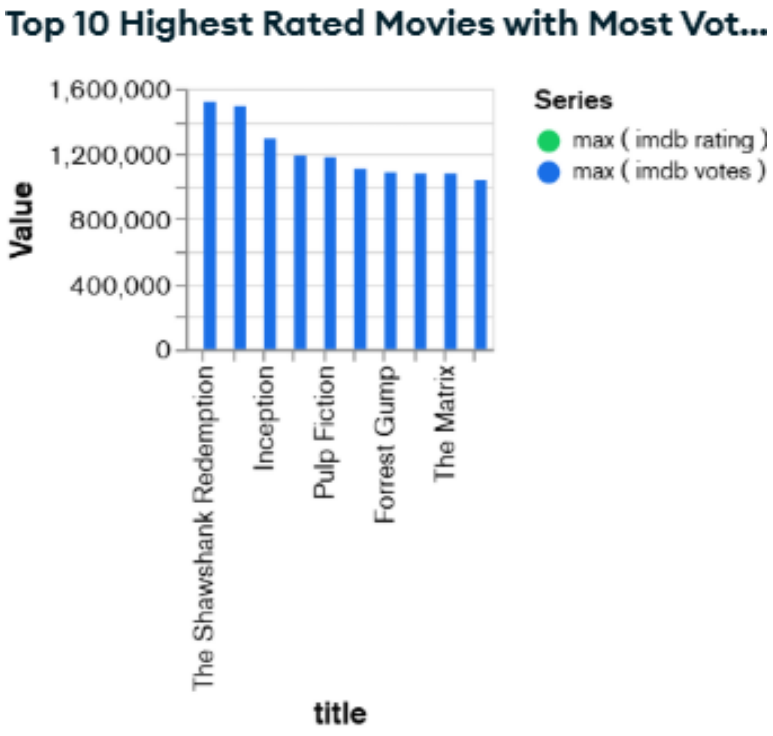
5.5.15. Share of Top 10 Most Popular Languages in Movies

- **English** dominates the dataset, implying that most of the movies in this collection are from English-speaking countries.
- **French, Spanish, and German** have a noticeable share, showing the global presence of European cinema.
- **Hindi and Mandarin** represent the influence of Bollywood and Chinese cinema, but their share is smaller compared to English-language films.
- Other languages such as **Japanese and Russian** also have a presence, highlighting international diversity in filmmaking.



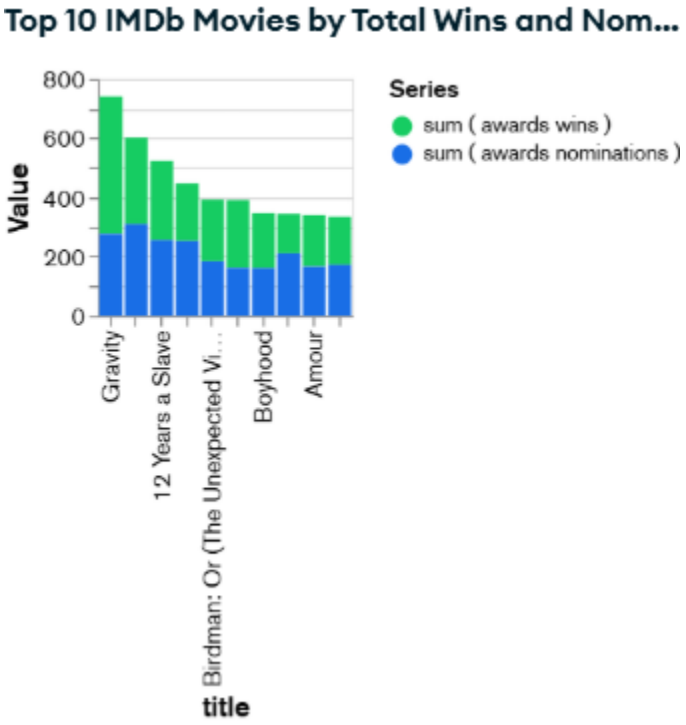
5.5.16. Top 10 Highest Rated Movies with Most Votes on IMDb

- **The Shawshank Redemption** has the highest votes and rating, reinforcing its status as one of the most beloved films.
- **Inception, Pulp Fiction, Forrest Gump, and The Matrix** also have high votes, indicating their widespread popularity and critical acclaim.
- The high number of votes suggests that these movies have a **strong fan base** and remain culturally significant.
- Most of these films belong to the **Drama, Thriller, or Sci-Fi** genres, suggesting these genres resonate well with IMDb users.



5.5.17. Top 10 IMDb Movies by Total Wins and Nominations

- **Gravity** has the highest combined wins and nominations, indicating strong critical and award recognition.
- **12 Years a Slave** and **Birdman** also have high award counts, suggesting the dominance of drama-based and socially relevant films in prestigious awards.
- The movies in this list generally have a **strong critical reception**, reinforcing the correlation between award recognition and quality filmmaking.
- A trend can be observed where **movies with strong narratives and technical excellence** receive the most awards.



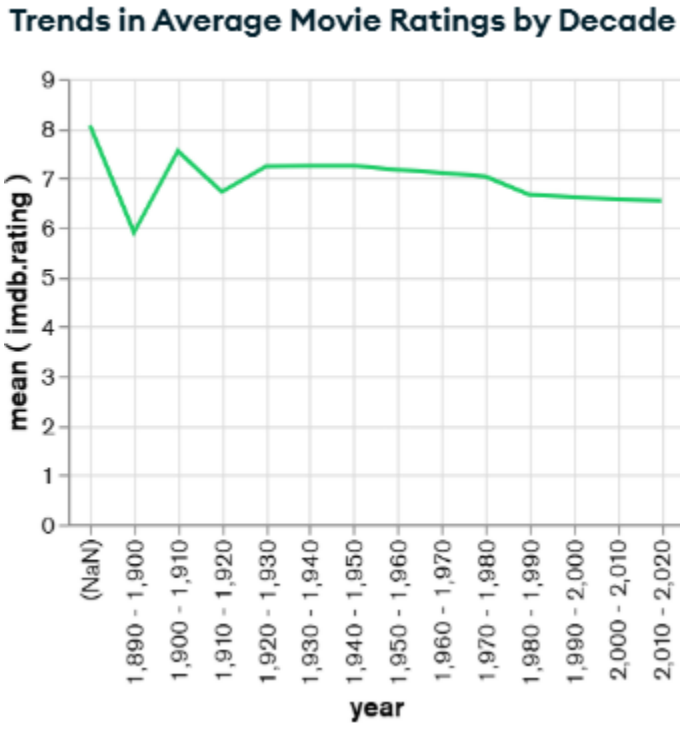
5.5.18. Top 10 Movies with the Highest IMDb Rating

- The movies listed have ratings between **5.5 and 7.6**, which suggests that they are not the highest-rated films overall but still hold moderate appreciation.
- **Édipussi (7.6) and ¡Vivan las Antípodas! (7.1)** are among the highest-rated, showing audience preference for unique storytelling or cultural significance.
- Some of the movies on this list have relatively **lower IMDb ratings** compared to what might be expected in a "Top 10" list, possibly indicating a **niche appeal** rather than mainstream popularity.

title	max (imdb.rat...
èxtasis	6.9
èvocateur: The Morton Downey Jr. Movie	6.8
èrido Movie	6.7
èon Flux	5.5
èlisa	6.7
èkte vare	6.2
èke och hans vèrld	6.7
èdipussi	7.6
èber uns das All	6.7
èVivan las Antipodas!	7.1
Total	**66.9**

5.5.19. Trends in Average Movie Ratings by Decade

- The average IMDb rating of movies has been **gradually declining** over time.
- Movies from the early 20th century (**1900-1950**) generally had **higher average ratings**, possibly due to fewer releases and a focus on high-quality productions.
- A noticeable decline starts from the **1980s onwards**, which may be attributed to an increase in mass production of films, changes in audience preferences, and the diversification of movie genres.
- Modern films (**2000s–2020s**) tend to have **lower ratings** compared to earlier decades, which could also reflect stricter audience critique and more widespread access to ratings.



5.5.20. Trends in Long Movies – Yearly Count of Films with Over 120 Minutes Runtime

- There has been a **steady increase** in the number of long movies (over 120 minutes) over the decades.
- From the **1980s onwards**, there is a **sharp rise**, showing a trend where filmmakers prefer longer runtimes, possibly to accommodate more complex storytelling.
- The peak around the **2000s–2010s** suggests a golden era for long-form cinema, with high-budget productions often exceeding 2 hours.
- A **slight dip** towards the most recent years might indicate a shift in audience preference towards shorter, more digestible content, influenced by streaming services and digital consumption habits.