CSCI 226 - Advanced Database Systems
Final Project Submission
Movie Database
Fall 2023

John Eagan, Department of Computer Science
California State University
Fresno, CA

Ritvik Gaur, Department of Computer Science
California State University
Fresno, CA

Rogelio Romero, Department of Computer Science
California State University
Fresno, CA

Anmol Singh, Department of Computer Science
California State University
Fresno, CA

November 2023

Submitted to Dr. David Ruby

Abstract

This database project centers on the development of a movie recommendation and information retrieval system using SQL scripts and database queries, emphasizing rapid access to real-time movie-related information. The primary objective involves crafting an efficient and comprehensive database schema housing diverse movie details, including genres, ratings, release dates, directors, actors, and user preferences.

Utilizing SQL queries, the system examines user behavior, preferences, and historical data to furnish tailored movie recommendations. Employing intricate SQL scripts and algorithms, the system strives to deliver precise and personalized movie suggestions to users, ensuring prompt access to pertinent movie details in real-time.

Problem Statement

The film industry generates vast amounts of data regarding movies, directors, actors, ratings, release dates, and more. Accessing and managing this data efficiently in real-time poses a significant challenge. To address this challenge, our project aims to design an SQL schema tailored for handling movie-related information effectively.

The problem revolves around the need for a robust database system capable of storing, organizing, and delivering movie data swiftly upon query. Key aspects of the problem include:

Data Complexity: Movies entail diverse information such as cast, crew, genres, release dates, ratings, reviews, and box office figures. Managing this complex and interconnected data in a structured manner is crucial.

Real-time Access: Immediate access to movie-related information is vital for various stakeholders, including viewers, critics, and industry professionals. A system that provides quick and seamless retrieval of relevant data is essential.

Scalability: As the film industry evolves and more movies get released, the database should be scalable to accommodate a growing volume of data without compromising performance.

Optimized Queries: The SQL schema should be designed to support optimized queries, allowing for efficient retrieval of specific movie details, director information, actor filmographies, ratings, and other related data points.

Our project aims to address these challenges by conceptualizing and implementing an SQL schema that efficiently organizes and manages movie-related data. This schema will provide a solid foundation for a database system capable of handling real-time queries and delivering accurate and up-to-date information about movies, directors, actors, and associated data.

Proposed Solution

To tackle the aforementioned challenges, our proposed solution involves the following key components:

Database Schema Design: We will meticulously design an SQL schema that efficiently organizes movie-related data. This schema will include well-structured tables to store information about movies, directors, actors, genres, ratings, reviews, release dates, box office figures, and more. Relationships between these entities will be established to ensure data integrity and ease of retrieval.

Normalization and Indexing: Employing normalization techniques, we will eliminate data redundancy and enhance data integrity. Additionally, strategic indexing will be implemented to optimize query performance, ensuring quick access to relevant information.

Scalability and Performance: Our schema design will consider scalability factors, allowing the database to handle increasing data volumes without compromising performance. This includes employing efficient storage mechanisms and partitioning strategies.

Query Optimization and Indexing Strategies: We will focus on creating optimized queries for frequently accessed data. By analyzing query patterns, we'll implement indexing strategies and query optimizations to enhance overall system performance.

Real-time Data Updates: Implementing mechanisms for real-time data updates to ensure that the database reflects the latest information about movies, ratings, reviews, and other relevant data points.

Our proposed solution aims to create a robust and scalable SQL schema specifically tailored for managing movie-related data. By focusing on efficient design principles, optimization strategies, and real-time updates, we anticipate delivering a database system capable of meeting the industry's demands for quick and accurate movie information.

Contents

1 Data Domain

1.1 Description

The movies domain encompasses a diverse range of audiovisual works created for the purpose of entertainment, cultural, educational, and artistic purposes. Movies are widely distributed through different forms of both physical and digital media in theaters, television, and streaming platforms across the world. The movie industry itself involves a multitude of tasks including production, distribution, exhibition, marketing, and critical evaluation, which enriches the domain with a vast, diverse collection of data from various sources.

1.2 Attributes

This section provides a description of the four categories of data that will comprise the majority of the database as well as their importance within the movies domain. A full description of the database attributes for each of the relations is instead provided in our database design diagram.

1.2.1 Movie Data

Movie-related data is the most abundant and clear type of data required for a movie database in order to provide comprehensive information relating to the fundamental details, descriptions, and characteristics of each film. These attributes remain mostly unchanged once a movie has been fully released and will serve as the primary "link" between the other relations in our database with the use of a unique identifier for each movie.

1.2.2 Cast & Crew Information

Movies are collaborative works of art and the individuals involved in their creation play pivotal roles in bringing these stories to wider audiences. In a movie-related database, the inclusion of detailed cast and crew information is essential for understanding the dynamics of filmmaking and giving context to the impact individuals made on the final film.

The cast members are the faces and voices that embody the characters in a film. Each actor brings a unique skill set, style, and interpretation to their roles. Detailed cast information, including character names, order of appearance, and associated details, preserves the artistic contributions of actors. Filmmakers can use this data to analyze the performance history of actors, assess casting decisions, and explore the evolution of characters across different projects.

Behind the scenes, the crew comprises a diverse group of individuals contributing their specialized skills to various departments. From directors and cinematographers to editors and costume designers, the crew shapes the visual and auditory elements of a movie. In this database, crew information provides a comprehensive view of the technical and creative expertise involved in filmmaking.

1.2.3 Critic & Audience Reviews

Movies are ultimately created with the intention of being viewed, appreciated, and critiqued by general audiences and film critics. The subjective nature of watching and experiencing a movie allows for different people to form different opinions about the film, and this diversity of perspectives provides further depth to the overall landscape of movie data.

Critic reviews, often written by seasoned experts in film, provide insightful analysis that can influence public perception, contribute to the discourse on filmmaking, and guide viewers in their movie choices. These reviews and critical ratings offer a qualitative assessment that goes beyond mere entertainment value, delving into aspects such as storytelling, cinematography, and performances.

On the other hand, audience reviews and ratings reflect the collective sentiment of everyday moviegoers. These opinions offer a democratic and inclusive perspective, capturing the tastes and preferences of a broad spectrum of viewers. Audience feedback can significantly impact a movie's success, as positive reviews may attract more viewers, while negative ones could influence potential audiences to explore alternative options.

1.2.4 Production, Budget, & Box Office Revenue

Understanding the financial aspects of movie production is essential for filmmakers, studios, and industry stakeholders. The budget allocated to a movie reflects the resources invested in its creation, including expenses related to cast salaries, special effects, set design, etc. Tracking this data provides insights into the scale and ambition of a project, enabling filmmakers to make informed decisions about resource allocation and production strategies while also informing audiences of the scope and potential of an upcoming film.

Box-office performance is a key indicator of a movie's success and its reception by the public. Revenue figures highlight the economic impact of a film, showcasing its ability to attract audiences and generate profits.

Analyzing box-office data allows industry professionals to assess market trends, identify successful genres or franchises, and make strategic decisions for future projects.

The involvement of production companies is a critical aspect of the movie-making process. Tracking production companies in a database provides a comprehensive view of the industry landscape. It helps identify key players, understand their contributions to successful films, and assess the global reach of their productions.

1.3 Data Sources

1.3.1 The Movies Dataset (Kaggle)

Description: Metadata on over 45,000 movies, including details such as titles, genres, and release dates. Additionally, the dataset comprises 26 million ratings provided by over 270,000 users.

1.3.2 Top 1000 Highest Grossing Movies (Kaggle)

Description: Compilation of data on the top 1000 highest-grossing Hollywood movies. This dataset provides insights into the financial success of these movies.

1.3.3 IMDB Top 250 Movies Dataset (Kaggle)

Description: A scraped dataset containing information about the most highly rated movies on IMDB. Specifically, it includes details on the top 250 movies as per IMDB ratings.

2 Tools/Technology

We are using Microsoft SQL Server 2022 developer edition for a database management system, SQL Server Management Studio (SSMS) for a development environment. Lucid chart for ER/UML diagrams and Transact-SQL (T-SQL) for the Query language.

2.1 SQL Server

Overview: SQL Server 2022 Developer Edition is a comprehensive database management system that plays a crucial role in our project. It is designed for development and testing purposes, providing a robust platform for building, testing, and deploying database solutions.

Key Features: Advanced Database Management: SQL Server 2022 Developer Edition offers a range of advanced features for efficient database management. These include support for high-performance transactions, in-memory processing, and advanced analytics.

Development and Testing Environment: As a dedicated developer edition, it provides an ideal environment for designing and testing database solutions. Developers can leverage its features without the constraints of production licensing.

Enhanced Security Measures: The Developer Edition includes robust security measures to protect sensitive data. This ensures that our development and testing environments maintain the same level of security as the production environment.

Compatibility and Integration: SQL Server 2022 Developer Edition seamlessly integrates with development tools, allowing for a smooth workflow. It ensures compatibility with other Microsoft technologies and supports various programming languages.

Use in Our Project: In our project, SQL Server 2022 Developer Edition serves as the primary database management system. It facilitates the following:

Database Design and Modeling: Enables the design and modeling of the database schema using SQL Server Management Studio (SSMS).
Query Development: Supports Transact-SQL (T-SQL) for developing queries and stored procedures.
Testing Environment: Provides a secure and reliable testing environment for ensuring the functionality and performance of our database solutions.
Collaboration: Facilitates collaboration among developers by allowing seamless sharing and version control of database scripts.

2.2 SQL Server Management Studio

SQL Server Management Studio (SSMS) is a graphical user interface (GUI) tool developed by Microsoft for managing and interacting with Microsoft SQL Server. It is closely related to SQL Server and serves as the primary interface through which database administrators, developers, and other users can interact with SQL Server databases. Here's how SSMS is related to SQL Server:

Interface for SQL Server Interaction: SSMS as the Tool: SSMS is the dedicated tool provided by Microsoft for interacting with SQL Server. It allows users to perform various tasks such as writing and executing queries, designing and modifying database schemas, managing security, monitoring performance, and more.

Direct Connection to SQL Server: SSMS establishes a direct connection to the SQL Server instance, providing a real-time and interactive environment for managing and working with databases.

Database Development and Administration: Development Tasks: SSMS supports database development tasks, such as creating, modifying, and deleting database objects (tables, views, stored procedures, etc.), writing and debugging Transact-SQL (T-SQL) queries, and managing database scripts.
Administration Tasks: For database administrators, SSMS provides tools for configuring server settings, managing security, monitoring server and database performance, and performing routine maintenance tasks.

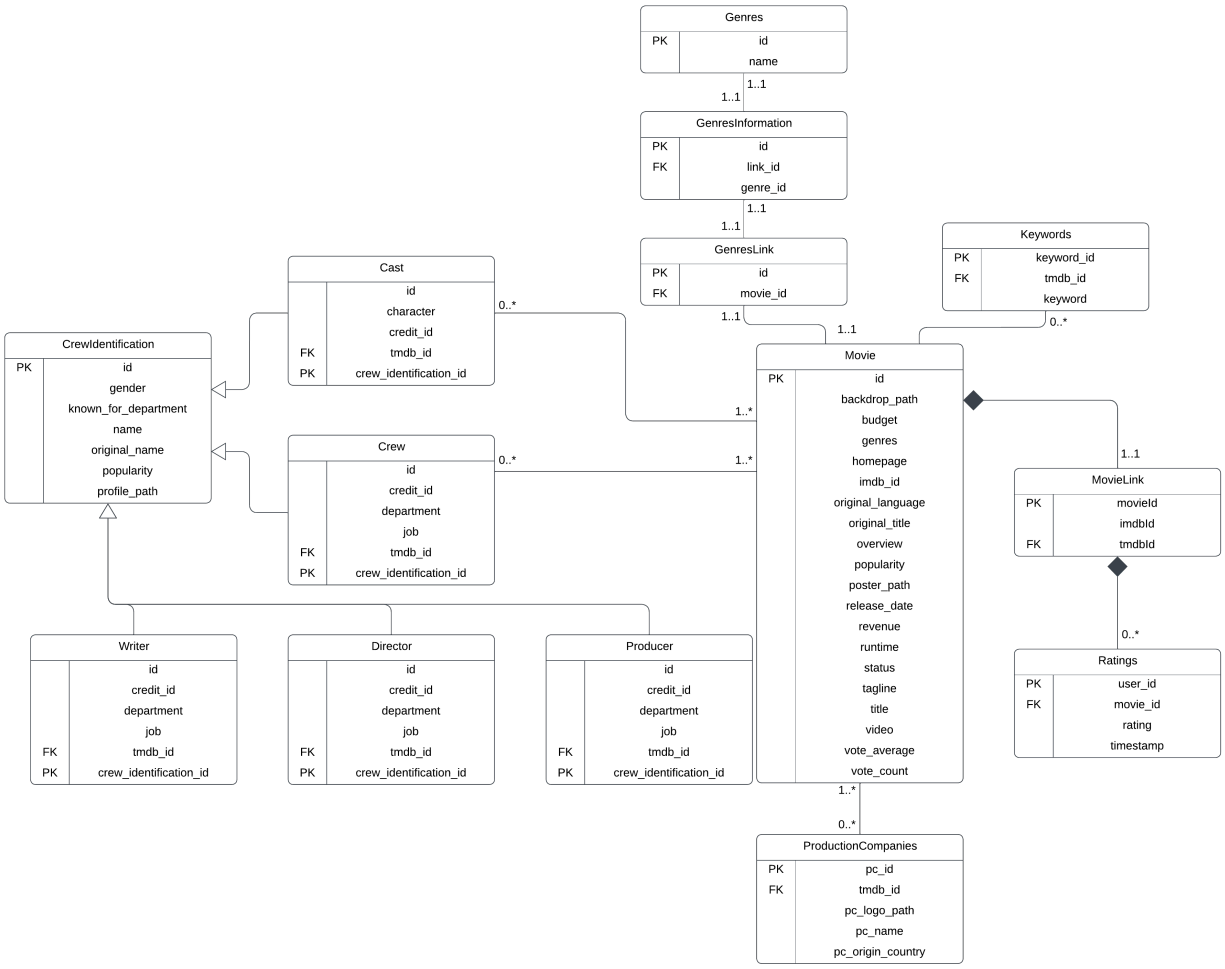
Integration with Other SQL Server Components: SSIS, SSRS, and SSAS Integration: SSMS seamlessly integrates with other components of the SQL Server ecosystem, including SQL Server Integration Services (SSIS), SQL Server Reporting Services (SSRS), and SQL Server Analysis Services (SSAS). This integration allows users to manage and monitor these services from a centralized interface.

Query Execution and Optimization: Query Editor: SSMS includes a powerful Query Editor that allows users to write and execute T-SQL queries. It provides features like syntax highlighting, code completion, and execution plan visualization for query optimization.

Security and User Management: Security Configuration: SSMS includes tools for configuring security settings, managing user roles, and setting permissions at various levels within the SQL Server instance.

3 Database Design

3.1 Design Diagram



4 Database Schema

The files required to define, create, and load our database are provided on Kaggle within Appendix B. SQL files related to triggers and stored procedures are also provided there.

4.1 Triggers

Triggers were used to create audit tables for each of the non-static tables in the database. Delete and update triggers were created that insert deleted data into the audit tables corresponding to the dbo table from which the data is being deleted. This allows us to keep an audit trail to see how the data has changed over time, and see when changes were made.

4.2 Stored Procedures

Several stored procedures were created to provide ease of use for common queries. The following stored procedures were created as part of the database schema:

- SearchMoviesByGenre : Retrieves movies associated with a genre. Takes GenreName as input.
- GetMoviesByReleaseYear : Retrieves movies that were released in the specified year. Takes ReleaseYear as input.
- GetMoviesByCrewMember : Retrieves movies that were worked on by a certain crew member. Takes CrewIdentificationID as input.

- GetMovieCastAndCrew : Retrieves the cast and crew that worked on a movie. Takes MovieID as input.
- GetMoviesByDirector : Retrieves the movies made by a specific director. Takes DirectorName as input.

5 Normal Forms

5.1 Functional & Multi-Valued Dependencies

The functional and multi-valued dependencies within the database relations are identified below. Note that all functional dependencies are considered to also be multi-valued dependencies by default.

Movies Metadata

- id → all other attributes
- title, release date → all other attributes

Crew Identification

- id → all other attributes
- name, profile path → all other attributes

Production Companies

- pc id, tmdb id → all other attributes
- pc id → pc logo path, pc name, pc origin country

Ratings

- user id, movie id → all other attributes
- user id, timestamp → movie id, rating

5.2 3rd Normal Form

There exist three violations of 3NF in the current database schema relating to:

1. pc id → pc logo path, pc name, pc origin country

The production companies relation utilizes a composite key of both the pc_id and tmdb_id to uniquely identify all the other attributes. This results in a violation of 3NF because the pc_id attribute itself is not a superkey and none of the production company-related data (logo path, name, origin country) are prime attributes.

An example of when this violation could result in database inconsistencies is when a production company changes their logo path, name, and/or origin country. If the update were to only occur on some subset of the movies (defined by tmdb_id), an inconsistency would develop between production companies with the same id and different logo path, name, or origin country.

5.3 Boyce-Codd Normal Form

Besides the previous set of violations of 3NF, there doesn't exist any further violations of BCNF. All other functional dependencies contain superkeys on the left-hand side, which is confirmed by the fact that they all functionally determine the rest of the attributes in their respective relations.

The violations cause unnecessary redundancy within the database because a production company and its associated data must be stored multiple times, once for each movie the company produces.

5.4 4th Normal Form

Once again, the previous violations result in violations of the 4NF. As there are no implicitly defined MVDs outside of the original FDs, all MVDs aside from pc id ↠ pc logo path, pc name, pc origin country contain a superkey on their left-hand side and do not result in any 4NF violations.

6 Queries & Analysis

The queries performed on this database can be found in Appendix C.

Most Common Keywords for Top 1000 Movies by Revenue

```
SELECT TOP 20 keyword, count(keyword) as number_of_occurences
FROM Keywords
WHERE tmdb_id in (SELECT top 1000 MovieMetadata.id
                  FROM MovieMetadata
                  ORDER BY revenue desc)
GROUP BY keyword
ORDER BY count(keyword) desc;
```

	keyword	number_of_occurences
1	based on novel or book	133
2	duringcreditsstinger	133
3	sequel	95
4	aftercreditsstinger	85
5	based on comic	83
6	superhero	69
7	new york city	62
8	dystopia	51
9	friendship	49
10	revenge	46
11	loss of loved one	41
12	based on young adult novel	39
13	alien	38
14	woman director	37
15	remake	37
16	musical	36
17	los angeles, california	36
18	magic	35
19	parent child relationship	34
20	anthropomorphism	34

Analysis

Evaluating what kinds of keywords are associated with movies that perform well in box-office revenue could provide important data about the kinds of topics and stories that general audiences would like to see. The results indicate that some of the most common keywords for successful movies are "based on novel or book", "during/after credits stringer", "sequel", "based on comic", "superhero", etc.

This indicates that most successful films are often based on an already established story from novels, books, or comics with the potential for sequels. This isn't particularly surprising given that established stories reduce the creative requirements of writing a new story and the relatively recent trend of large-scale superhero-based film collections. Further refinement of this query could isolate the movies to particular ranges of release dates in order to see how keywords evolve over time and what trends have traditionally led to box office success in the movie industry.

Highly Contentious Movies

```
SELECT oneTable.id, oneTable.title, ones, fives, ones+fives as total_ones_and_fives
FROM (SELECT id, title, count(rating) as ones
      FROM MovieMetadata, Ratings, MovieLink
      WHERE Ratings.movie_id = MovieLink.movieId and MovieMetadata.id = MovieLink.tmdbId
      ↪  and rating = 1
      GROUP BY id, title) as oneTable,
     (SELECT id, title, count(rating) as fives
      FROM MovieMetadata, Ratings, MovieLink
      WHERE Ratings.movie_id = MovieLink.movieId and MovieMetadata.id = MovieLink.tmdbId
      ↪  and rating = 5
      GROUP BY id, title) as fiveTable
WHERE oneTable.id = fiveTable.id and ones+fives > 100 and abs(fives*1.0/ones - 1) < 0.25
ORDER BY total_ones_and_fives desc;
```

	id	title	ones	fives	total_ones_and_fives
1	3049	Ace Ventura: Pet Detective	4250	3338	7588
2	8467	Dumb and Dumber	4281	3270	7551
3	771	Home Alone	1657	2069	3726
4	95	Armageddon	1770	1825	3595
5	9350	Cliffhanger	1287	1596	2883
6	75	Mars Attacks!	1573	1195	2768
7	9739	Demolition Man	1019	1132	2151
8	2758	Addams Family Values	1130	964	2094
9	955	Mission: Impossible II	1135	948	2083
10	8005	Robin Hood: Men in Tights	1025	1023	2048

Analysis

Contentious movies are defined as movies with a large amount of disagreement in the general audience's rating of the movie. For the sake of this analysis, contentious movies will be considered as movies with a large number of both 1 and 5 star ratings, implying that moviegoers either found this movie to be horrible or a cinema masterpiece. Specifically, the query requires the movies to have a ratio of 1 to 5 star ratings that's within 0.75 to 1.25.

Based on the results, there seems to be a large amount of contention surrounding movies with a large amount of violent or pronounced behavior for the sake of comedic effect, including movies like Ace Venture, Dumb and Dumber, Home Alone, etc.

Genres with the Highest Average Popularity

```sql
SELECT g.name, AVG(ci.popularity) AS AveragePopularity
FROM Genres g
JOIN GenresInformation gi ON g.id = gi.genre_id
JOIN GenresLink gl ON gi.link_id = gl.id
JOIN MovieMetadata m ON gl.movie_id = m.id
JOIN CrewIdentification ci ON m.id = ci.id
GROUP BY g.name
ORDER BY AveragePopularity DESC;
```

	name	AveragePopularity
1	Family	2.18216983972437
2	War	2.15728989440643
3	Music	2.12741065016093
4	TV Movie	2.12268389148145
5	Animation	2.09976439175033
6	Thriller	2.0674066490861
7	Horror	2.06190429758052
8	Comedy	2.0544038472018
9	Fantasy	2.05117697482978
10	Adventure	2.02929716848448
11	History	2.02723669462655
12	Crime	2.02687634550681
13	Drama	2.02425288339719
14	Action	2.02251871502484
15	Science Fiction	2.02150536614631
16	Romance	2.0114601078333
17	Documentary	1.99681654403042
18	Mystery	1.98315120473817
19	Western	1.86038546919823

Analysis

Understanding the popularity of genres is valuable for filmmakers, studios, producers, and distributors. It can provide insight into the preferences of audiences, providing a metric that is a major contributing factor to a movie's popularity.

The results of this query indicate that family movies and war movies are the most popular movies amongst the entries in the database, and western and mystery movies rank as the most unpopular. Family movies likely appeal to a wide audience, being enjoyed by children, teenagers, and adults. They typically focus on positive and heartwarming themes, that provide the audience with a feel-good experience. They are also likely enjoyed by families together, leading to higher viewership and therefore, higher ratings.

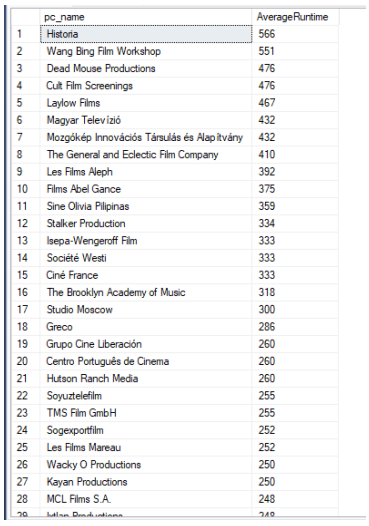
War movies often depict significant historical events, attracting audiences with an interest in history. They often feature intense storytelling and set pieces, that can create compelling narratives that resonate and captivate audiences.

The least popular movies likely appeal to a smaller and older audience. Western and mystery movies are often slower in pace, which can be seen as boring by younger audiences. The popularity of Western movies has likely also been affected by changing interests over time.

This query can be interesting in exploring other correlations between data, such as if there is a correlation between popularity and other factors, such as budget or release dates.

Average runtime of movies produced by each production company in descending order

```
SELECT g.name, AVG(ci.popularity) AS AveragePopularity
FROM Genres g
JOIN GenresInformation gi ON g.id = gi.genre_id
JOIN GenresLink gl ON gi.link_id = gl.id
JOIN MovieMetadata m ON gl.movie_id = m.id
JOIN CrewIdentification ci ON m.id = ci.id
GROUP BY g.name
ORDER BY AveragePopularity DESC;
```



	pc_name	AverageRuntime
1	Historia	566
2	Wang Bing Film Workshop	551
3	Dead Mouse Productions	476
4	Cult Film Screenings	476
5	Laylow Films	467
6	Magyar Televízió	432
7	Mozgókép Innovációs Társulás és Alapítvány	432
8	The General and Eclectic Film Company	410
9	Les Films Aleph	392
10	Films Abel Gance	375
11	Sine Olivia Pilipinas	359
12	Stalker Production	334
13	Isepa-Wengeroff Film	333
14	Société Westi	333
15	Ciné France	333
16	The Brooklyn Academy of Music	318
17	Studio Moscow	300
18	Greco	286
19	Grupo Cine Liberación	260
20	Centro Português de Cinema	260
21	Hutson Ranch Media	260
22	Soyuztelefilm	255
23	TMS Film GmbH	255
24	Sogexportfilm	252
25	Les Films Mareau	252
26	Wacky O Productions	250
27	Kayan Productions	250
28	MCL Films S.A.	248

Analysis

The runtime of a movie can differ by the genre of the movie. Family films will typically be shorter in length as their primary target audience is children. Historical films, such as documentaries, will typically be longer in length, due to their expositional nature and focus on details and storytelling.

The top several production companies with the longest average runtimes produced historical films. The content is meant to be consumed over time, including films that are broken down into mini-series episodes.

Most of the companies with the shortest runtimes produced family-oriented content. These companies include the Walt Disney Company and Warner Brothers. The average runtimes of these production companies fall under the half-hour range, due to their production of short-form content, such as cartoons.

This metric can be used to analyze whether the popularity and revenue of a production company is significantly impacted by the runtime of the content that they are producing. It has the potential to be used for optimality purposes, helping to determine if a company is more profitable producing short-form content or if they benefit from longer films.

Comparing movies released on the same day and comparing which had more budget.

```
WITH MoviePairs AS (
SELECT
    m1.title AS movie1,
    m2.title AS movie2,
    m1.release_date,
    m1.revenue AS revenue_movie1,
```

```
        m2.revenue AS revenue_movie2
    FROM
        MovieMetadata m1
        JOIN MovieMetadata m2 ON m1.release_date = m2.release_date AND m1.id < m2.id
)
SELECT
    movie1,
    movie2,
    release_date,
    CASE
        WHEN revenue_movie1 > revenue_movie2 THEN movie1
        WHEN revenue_movie2 > revenue_movie1 THEN movie2
        ELSE 'Both movies have similar revenue'
    END AS more_cost_to_make
FROM
    MoviePairs;
```

Results | Messages

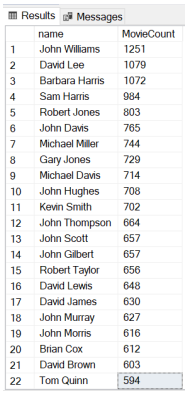
	movie1	movie2	release_date	more_cost_to_make
1	Die Hard	A Fish Called Wanda	1988-07-15	Die Hard
2	3-Iron	The Dust Factory	2004-10-15	3-Iron
3	Blood Diamond	The Holiday	2006-12-08	The Holiday
4	Talk to Her	Resident Evil	2002-03-15	Resident Evil
5	Together	Bring It On	2000-08-25	Bring It On
6	The Wrong Trousers	What's Eating Gilbert Grape	1993-12-17	What's Eating Gilbert Grape
7	Mr. Bean's Holiday	Reign Over Me	2007-03-22	Mr. Bean's Holiday
8	Valkyrie	Buddenbrooks	2008-12-25	Valkyrie
9	Rocky Balboa	Night at the Museum	2006-12-20	Night at the Museum
10	Ariel	The Bear	1988-10-21	The Bear
11	Resident Evil: Apocalypse	Land of Plenty	2004-09-10	Resident Evil: Apocalypse
12	Requiem for a Dream	Meet the Parents	2000-10-06	Meet the Parents
13	Resident Evil: Apocalypse	I ♥ Huckabees	2004-09-10	Resident Evil: Apocalypse
14	Land of Plenty	I ♥ Huckabees	2004-09-10	Both movies have similar r...
15	Ocean's Eleven	The 51st State	2001-12-07	Ocean's Eleven
16	Minority Report	Erkan & Stefan 2	2002-06-20	Minority Report
17	American History X	Armageddon	1998-07-01	Armageddon
18	Together	There's Only One Jimmy G...	2000-08-25	Both movies have similar r...
19	Bring It On	There's Only One Jimmy G...	2000-08-25	Bring It On

Analysis

There have been many recorded dates where there is release of two similar movies on the same date due to which there is confusion among people to select which movie to watch which have direct effect on the income of the movie. So using this query we can simply see the budget of the movie in which it was made and assume the revenue loss of movie with higher budget as it got clashed with the movie of lower budget. Using the metrices from the past data one can easily evaluate that is it beneficial to release a movie on the same date to overlap the other movie or not, based on the resources spent.

Crew members worked in most number of movies

```
SELECT ci.name, COUNT(cm.tmdb_id) AS MovieCount
FROM CrewIdentification ci
JOIN (SELECT DISTINCT tmdb_id, crew_identification_id FROM Cast
      UNION ALL
      SELECT DISTINCT tmdb_id, crew_identification_id FROM Crew
      UNION ALL
      SELECT DISTINCT tmdb_id, crew_identification_id FROM Director
      UNION ALL
      SELECT DISTINCT tmdb_id, crew_identification_id FROM Producer
      UNION ALL
      SELECT DISTINCT tmdb_id, crew_identification_id FROM Writer) cm
ON ci.id = cm.crew_identification_id
GROUP BY ci.name
ORDER BY MovieCount DESC
```



Results | Messages

	name	MovieCount
1	John Williams	1251
2	David Lee	1079
3	Barbara Harris	1072
4	Sam Harris	984
5	Robert Jones	803
6	John Davis	765
7	Michael Miller	744
8	Gary Jones	729
9	Michael Davis	714
10	John Hughes	708
11	Kevin Smith	702
12	John Thompson	664
13	John Scott	657
14	John Gilbert	657
15	Robert Taylor	656
16	David Lewis	648
17	David James	630
18	John Murray	627
19	John Morris	616
20	Brian Cox	612
21	David Brown	603
22	Tom Quinn	594

Analysis

The SQL query you provided aims to gather information about the count of movies associated with each crew member across various roles such as cast, crew, director, producer, and writer. It can help us to see the most popular faces in crew as they are working in most of the movies due to their talent/popularity or hard work that helps movie to be in limelight also helps in selecting the best crew based on data for producing a movie.

7 Visualizations

- **Genre-Keyword Clouds:**
  Visit Genre-Keyword Clouds to view this visualization.
- **Budget, Revenue, and Profit**
  Visit Top Movies By Revenue to view this visualization.
- **Critic Reviews/Ratings**
  Visit Critic Reviews to view this visualization.
- **Production Companies Around the Globe**
  Visit Production Companies Map to view this visualization.
- **Popularity of Movie Casts**
  Visit Movie Casts Popularity to view this visualization.

A Datasets

1. TMDB Movies Dataset
2. Top 1000 Highest Grossing Movies
3. Top 250 Movies Dataset
4. Movies Combined & Cleaned Dataset

B Database Schema Files

Database Creation

C Query Files

Database Queries