**End-to-End Data Engineering Project: Analyzing IMDB Movie Data**

**Tools: Alteryx, SQL Server Management Studio, SQL, Power BI**

* **By Santhi Kiran Chavali**

**Introduction:**

Nowadays, decision-making in many businesses is based on data, thus being able to extract, transform, and analyze big datasets is critical. This project showcases a comprehensive approach to data engineering and analysis using a strong toolkit of industry-standard technologies, all centered around the rich and complex world of Cinema.

The IMDB Movie Data Analysis project illustrates the end-to-end process of handling a huge dataset, from initial data collecting to insightful visualization. This project demonstrates the entire range of data engineering possibilities by utilizing a potent combination of Alteryx for data preparation, Microsoft SQL Server for data storage and sophisticated querying, and Power BI for dynamic display.

**The Dataset:**

The dataset, which was obtained via Kaggle, includes a plethora of movie-related data, which presents a complex, real-world scenario that calls for sophisticated data engineering techniques. It has 21 diverse columns, including 'imdbID', 'title', 'year', 'rating', 'runtime', 'genre','released', 'director', 'writer', 'cast','metacritic', 'imdbRating', 'imdbVotes', 'poster', 'plot', 'fullplot', 'language', 'country', 'awards', 'lastupdated', and 'type'.

**The Problem Statement:**

The project's goal is to provide insights into movie ratings, genres, and trends by analyzing an extensive IMDB movie dataset. The task at hand involves effectively cleansing, storing, and presenting this data to enable well-informed decision-making within the film industry.

**The Approach:**

* Clean and prepare raw data using Alteryx data manipulation tools
* Design a robust database in SSMS and import the cleaned data
* Perform Complex SQL Queries to extract meaningful insights from the data
* Create interactive visualizations using Power BI to present our findings

**Clean and prepare the data using Alteryx:**

1. **Data import:**

The IMDB dataset was successfully imported into the workflow once a connection was made using Alteryx's Input tool.

1. **Data Type Modification and Column Selection:**

* All columns' data types were changed from their original string format to more suitable kinds by using the Select tool.
* Furthermore, data like movie ID, non-IMDB rating, Metacritic score, poster, plot, full plot, and latest updated date were eliminated as they were deemed redundant and useless.
* The dataset was streamlined to concentrate on the most important data for analysis.

1. **Handling missing values:**

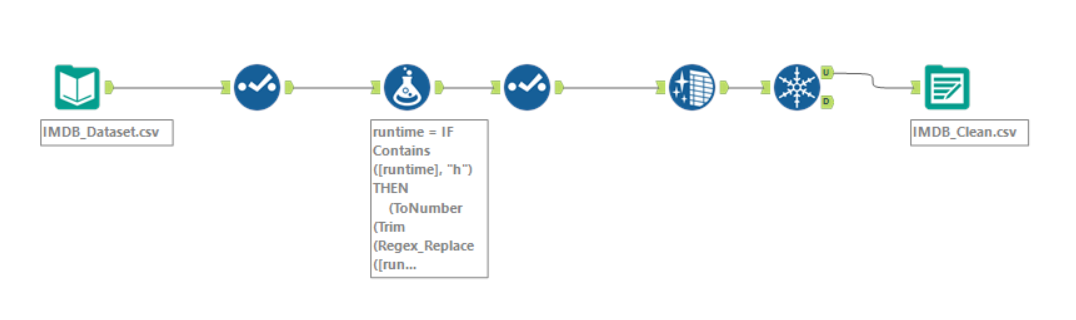
In columns like "released," "writer," and "language," missing data were noted. These entries were kept for a more thorough investigation rather than being deleted, recognizing that even imperfect data can yield insightful information.

1. **Runtime Standardization:**

To standardize the 'runtime' column, several time formats (such as "min" or "1hr and min") were converted into a uniform numerical representation in minutes using the Formula tool.

1. **Data Cleaning, Deduplication, and Export:**

* To ensure consistency and enhance data quality, unnecessary spaces were eliminated from data fields using the Data Cleansing tool.
* After that, to preserve data integrity and avoid biased analytic results, duplicate entries in the dataset were found and eliminated using the Unique tool.
* The cleaned, deduplicated, and prepared dataset was then exported using the Output tool, completed the Alteryx workflow and preparing the data for further analysis.



**Creating Database and importing Data in SQL Server Management Studio:**

After the data preparation in Alteryx, The next important step was to create a strong database structure and import the cleaned dataset into SQL Server. This process was carried out smoothly using SQL Server Management Studio, displaying competence with database administration and data import strategies.

1. **Database creation:**

A new database was created in SSMS specifically for the IMDB data which involved defining the database name and setting up necessary parameters to provide optimal performance and storage

1. **Table Structure Design:**

A table structure was thoughtfully established inside the newly created database to hold the cleaned dataset. The table schema was created with the proper data types allocated to each column in order to match the structure of the Alteryx output. This phase needed a deep comprehension of SQL data types and how they affect query performance and data storage.

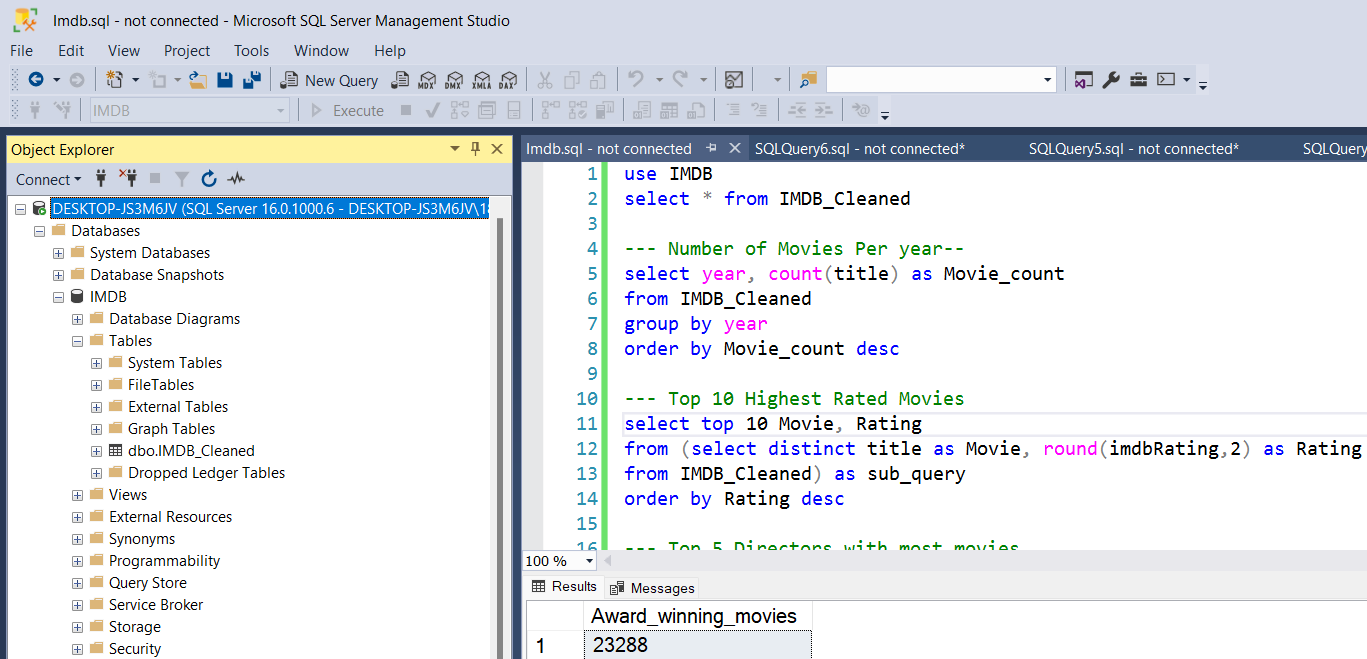
1. **Data Import Process:**

The SQL Server Import and Export Wizard was utilized to efficiently transfer the cleaned data from the Alteryx output file into the newly created table. This process involved:

* Selecting the source file (the Alteryx output)
* Mapping the source columns to the destination table columns
* Configuring any necessary data type conversions
* Executing the import process

1. **Data Verification:**

Following the import procedure, several SELECT queries were executed in order to confirm the accuracy and comprehensiveness of the imported data. This stage guaranteed that all records were transferred accurately and that the proper data types were maintained.



**SQL Data Analysis:**

Following the data's successful import into SQL Server, several SQL queries were run to glean insightful information from the IMDB movie dataset. These queries demonstrate a mastery of SQL and the capacity to extract useful data from complex datasets. These are a few of the major analyses that were done:

|  |  |
| --- | --- |
| **Movie Production Trend:** | **Top Rated Movies:** |
| **Prolific Directors:** | **Genre-based runtime:** |
| **Movie Production by Country:** | **IMDB Rating Trends:** |
| **Popular Actors:** | **Award-winning Movies:** |

**Data Visualization with Power BI:**

The final step of the project was to use Power BI to create dynamic and interactive representations after the SQL analysis. This phase showed mastery of data visualization techniques and the capacity to effectively communicate complicated data stories by transforming unprocessed data and SQL query results into visually attractive insights.

A screenshot of a computer

Description automatically generated

**Key Visualizations and Insights:**

1. **Number of Movies and Award-winning movies:** The dashboard shows there are 46,014 movies in total. 23,288 of these are classified as award-winning films, which makes up a sizeable chunk of the dataset (almost 50%).
2. **Top 10 Highest-Rated Movies:** The highest-rated films in the dataset are displayed in a table representation. Remarkably, "Band of Brothers" has the highest rating on IMDB, with an astounding 9.60. "Planet Earth" and "The Chaos Class" are among the few titles that are tied at 9.50. This representation makes it easy to find the database's most highly regarded material.
3. **Directors with Most Movies:** The most prolific directors in the dataset are displayed in a bar chart. Among the most productive are John Ford, Georges Méliès, and Michael Curtiz. The directors who have produced the most films and had the most effect on the film business are highlighted in this visualization.
4. **Movie count by year:** A line graph showing the number of movies made each year shows some intriguing historical trends in the cinema industry. The early 20th century saw a discernible rise in the number of movies produced, with notable peaks in a few years. Periods of intense film activity and prospective years of economic boom are indicated by this image.
5. **Percentage of Movies by Language:** With 73.13% of all movies, the English language is most prevalent, as seen by the pie chart. French (5.15%), Japanese (4.27%), Italian (3.72%), and Spanish (3.22%) are among the other noteworthy languages. The prevalence of English-language films is indicative of Hollywood's and other English-speaking nations' influence on world cinema.

**Key Learnings:**

1. **Data pipeline development:** This project demonstrated the importance of creating a robust data pipeline, from data cleaning in Alteryx to database management in SQL Server and visualization with Power BI.
2. **Data Cleaning Techniques:** The Alteryx application brought to light the importance of meticulous data preparation and cleaning to guarantee data quality for accurate analysis.
3. **SQL Proficiency:** Developing complex SQL queries enhanced skills in data extraction and manipulation which is crucial for deriving meaningful insights from large datasets
4. **Visualization:** The Power BI dashboard provided an example of how skillful visualization can convert unstructured data into useful insights and enable stakeholders to access complex information.

**Conclusion:**

This IMDB movie data analysis project effectively illustrated a thorough method of data engineering and analysis. The project demonstrated the ability to manage massive information, conduct in-depth research, and display findings in a visually appealing way by utilizing tools like Alteryx, SQL Server, and Power BI. Decision-makers in the film industry can benefit greatly from the insights gathered, which include the ability to identify highly rated films, filmmakers with the highest production values, and patterns in filmmaking throughout time. Specifically, the interactive Power BI dashboard offers an effective tool for dynamically investigating these insights.

With this project, I was able to get more insight into the film industry's landscape in addition to improving my technical abilities in data processing, analysis, and visualization. Complete data engineering capabilities are reflected in the process from data cleaning to final visualization, showing that the team is prepared to take on challenging data problems in the workplace.

Overall, this project functions as a powerful portfolio piece, demonstrating the capacity to extract significant insights from unstructured data and effectively communicate them, skills that are highly sought after in the data engineering and analytic community.