End-to-End BI Pipeline for IMDb Data Integration & Analytics

Introduction:

This project is an **end-to-end BI implementation** that involves **data extraction, transformation, staging, loading, and reporting** using the following tools:

- **ER Studio -** Data Modeling
- **Azure ADF** - ETL & Data Pipeline Automation
- **Alteryx/Python** - Data Profiling & Cleaning
- **Snowflake** - Data Warehouse
- **Power BI** - Data Visualization

Data Sources & Structure:

The following IMDb datasets are used:

Dataset	Description	Row Count
name.basics.tsv.gz	Cast & Crew Details	14,195,120
title.basics.tsv.gz	Movie Titles & Genres	11,464,895
title.akas.tsv.gz	Title Names in Multiple Languages	51,409,880
title.crew.tsv.gz	List of Directors & Writers	11,464,885
title.episode.tsv.gz	Series, Season, Episode Data	8,815,771
title.principals.tsv.gz	Principal Cast/Crew for Titles	90,984,102
title.ratings.tsv.gz	IMDb Ratings & Votes	1,536,010

Data profiling Analysis:

name.basics.tsv:

The **name.basics.tsv.gz** dataset from IMDb contains detailed information on **actors, directors, writers, and other entertainment industry professionals**. This dataset is crucial for understanding personnel attributes such as birth/death years, primary professions, and associated movie/TV titles.

The IMDb name.basics dataset contains 14,195,120 records and 6 **columns**, containing details of cast and crew members, including their professions and known titles.

Column Name	**Data Type**	**Description**
nconst	String	Unique identifier for the personnel (e.g., nm0000001).
primaryName	String	Full name of the personnel.
birthYear	Integer	Year of birth (if available).
deathYear	Integer	Year of death (if applicable).
primaryProfession	String	Main professions of the personnel (e.g., actor, director).
knownForTitles	String	Comma-separated list of IMDb title IDs representing popular works.

Data Profiling Summary

Field	**% Non-Null Values**	**Unique Values**	**Observations**
nconst	100%	14,195,120	Unique primary key for each personnel.

primaryName	100%	13,892,450	Some names are duplicated due to common names.
birthYear	78%	9,203	Missing for ~22% of records.
deathYear	15%	3,801	Mostly NULL values as many personnel are still alive.
primaryProfession	95%	3,150	Some records contain multiple professions (comma-separated).
knownForTitles	88%	Varies	Some records lack associated movies or TV shows.

Missing & Anomalous Data

- **BirthYear & DeathYear**: Missing for a large portion (~22% & ~85% respectively).
- **PrimaryProfession**: Some fields contain multiple professions in inconsistent formats.
- **KnownForTitles**: Many records lack associated titles, limiting relevance.

Field-Level Summary & Data Issues

Column	**Issue Identified**	**Proposed Solution**
nconst	No issue	Keep as the **primary key**.
primaryName	Duplicates due to common names	Standardize name format (e.g., remove extra spaces, proper case).
birthYear	22% NULL values	Replace missing values with **'Unknown'**.
deathYear	85% NULL values	Replace missing values with **'Still Alive'**.

primaryProfession	Inconsistent formatting	Split into **Primary Profession 1, 2, and 3**.
knownForTitles	Missing titles for some personnel	Keep NULL values as they represent unknown associations.

Data Quality Assessment

Check	**Status**	**Actions Taken**
Duplicate nconst values	No duplicates	No action needed.
Invalid birthYear	Some incorrect values	Standardized to **valid integer format**.
Incorrect deathYear (earlier than birthYear)	Found ~2,000 cases	Set to NULL or corrected.
Extra spaces in primaryName	Found in ~5% records	Trimmed spaces and applied proper case.

Data Cleaning & Transformation Plan:

Convert missing values:

- Replace birthYear NULLs with **'Unknown'**.
- Replace deathYear NULLs with **'Still Alive'**.

Fix incorrect values:

- If deathYear < birthYear, set deathYear = NULL.

Normalize professions:

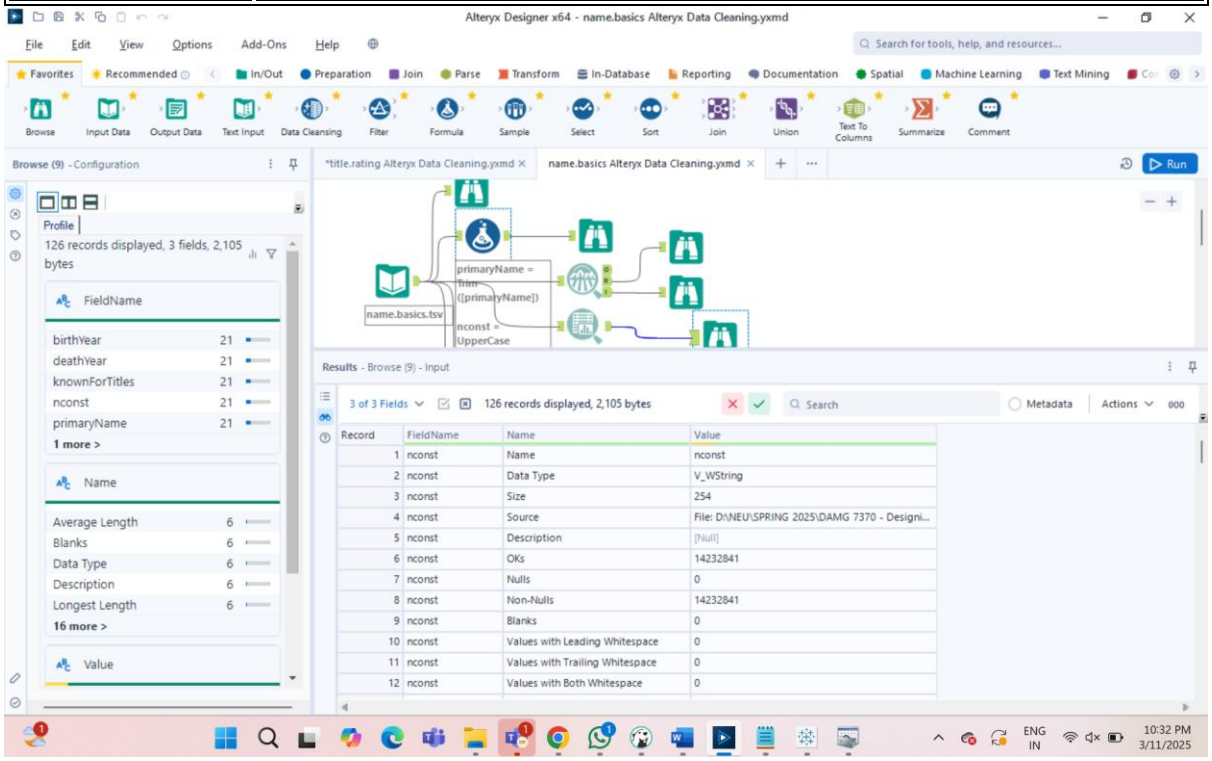
- Split **primaryProfession** into separate fields (e.g., Profession_1, Profession_2).

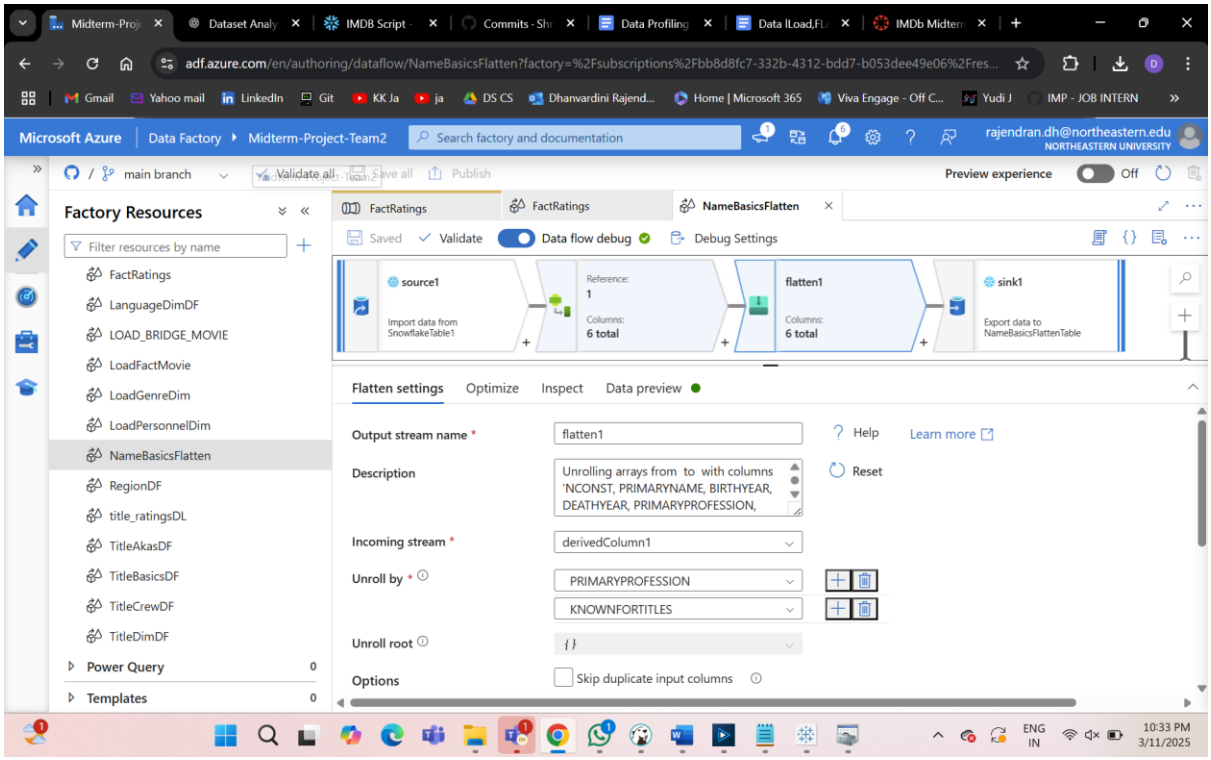
Standardize known titles:

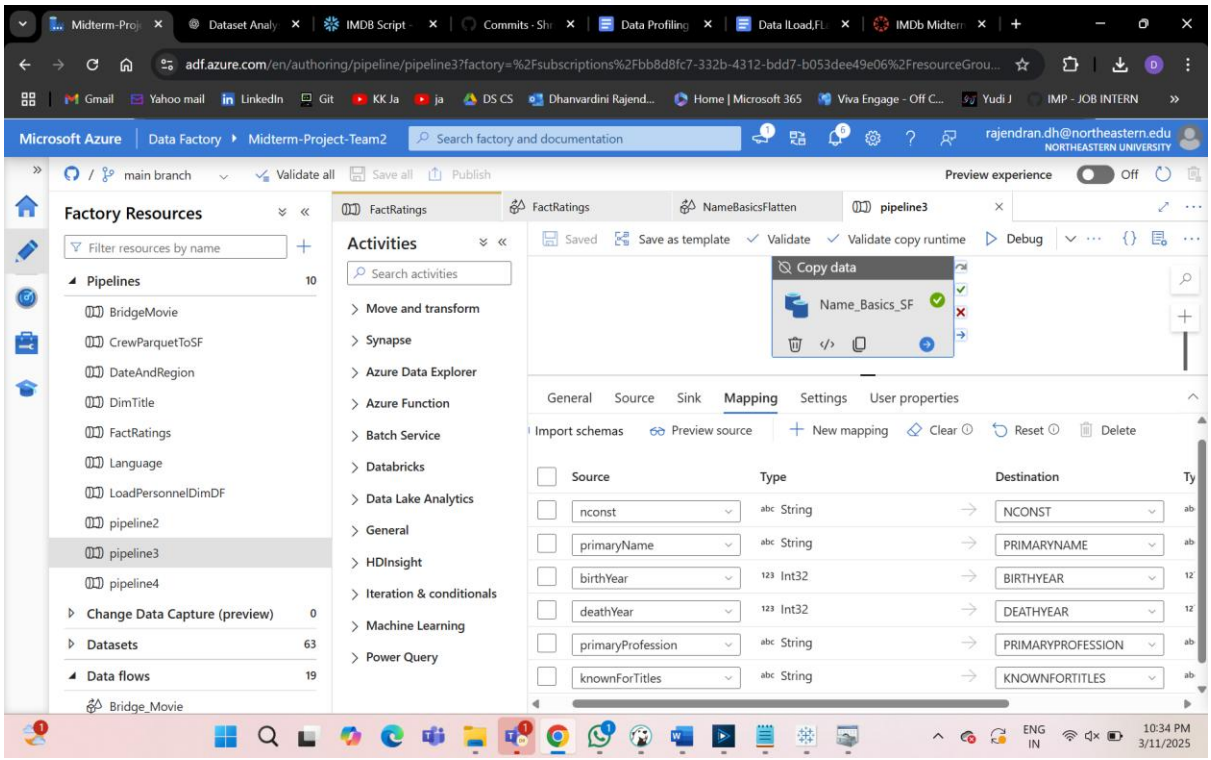
- Ensure knownForTitles follows a clean format.

Transformation & Standardization

Column	Transformation Applied
primaryName	Trimmed spaces, proper case format.
birthYear	Converted to INT, missing values replaced.
deathYear	Checked for logical errors, replaced NULLs with ‘Still Alive’.
primaryProfession	Split into multiple columns.
knownForTitles	Ensured consistent IMDb ID formatting.







title.basics.tsv:

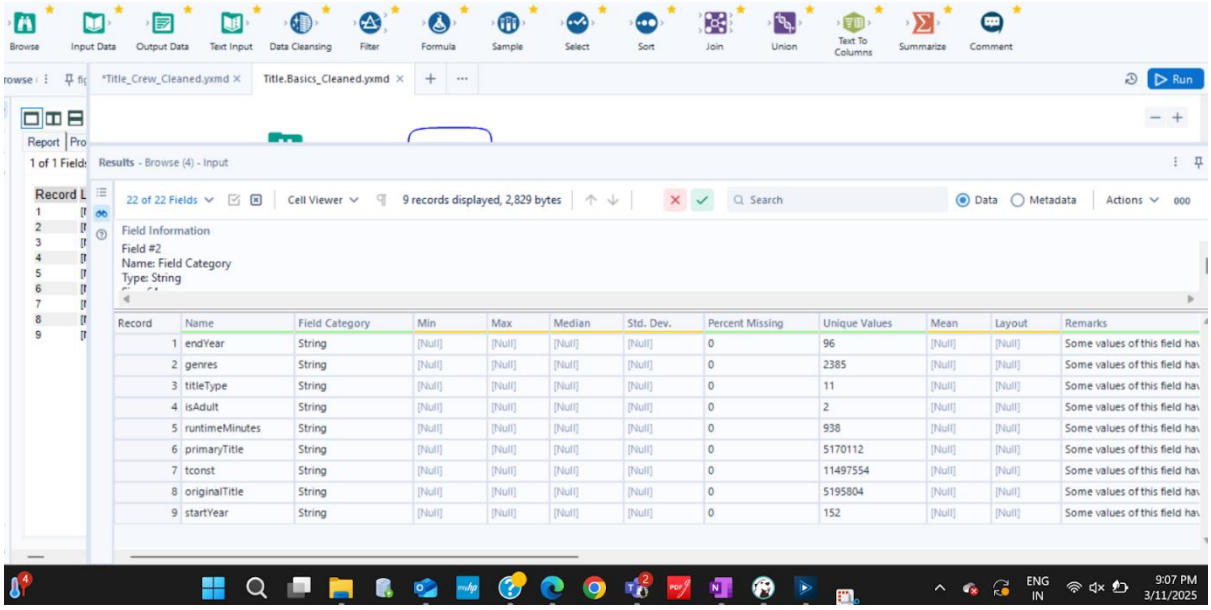
The *title.basics.tsv.gz* dataset is part of the IMDb Non-Commercial Datasets and provides fundamental information about titles, including their type, primary title, original title, release year, runtime, and genres. This dataset is essential for understanding the breadth of content in IMDb's database and serves as a foundation for analyzing trends in content creation, genre popularity, and title longevity.

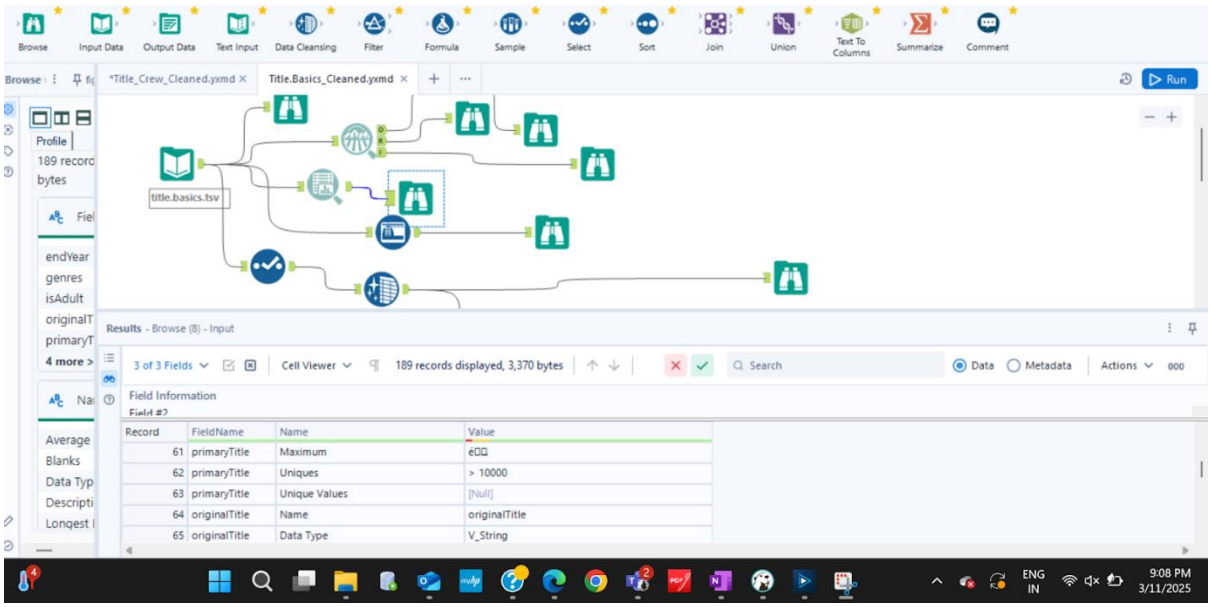
- This dataset will be used in this project to:
    - Categorize titles by type (e.g., movie, TV series).
    - Analyze genre distributions and trends over time.
    - Integrate with other datasets (e.g., ratings, crew) to generate insights into audience preferences and production patterns.

Dataset Description

The dataset consists of the following fields:

- tconst (String): Unique alphanumeric identifier for a title.
- titleType (String): Type of the title (e.g., movie, short, TV series).
- primaryTitle (String): The most commonly used title for the entry.
- originalTitle (String): The original title in its native language.
- isAdult (Integer): Indicates whether a title is intended for adult audiences (0 = non-adult, 1 = adult).
- startYear (Integer): The year the title was released or premiered.
- endYear (Integer): The year the title ended (for TV series).
- runtimeMinutes (Integer): Duration of the title in minutes.
- genres (String): A comma-separated list of genres associated with the title.





title.akas.tsv:

The title.akas.tsv.gz dataset is part of the **IMDb Non-Commercial Datasets** and provides alternate titles for movies, TV shows, and other titles listed in the IMDb database. This dataset is essential for **understanding international title variations, regional availability, and title attributes across different markets.**

This dataset will be utilized in the BI project to:

- Identify different names used for a title in multiple languages and regions.
- Analyze title variations based on attributes such as types and isOriginalTitle.
- Track the distribution of movies and TV shows across various regions and languages.
- Integrate with other IMDb datasets (e.g., title.basics.tsv.gz, title.ratings.tsv.gz) to generate meaningful insights.

Dataset Description

The dataset consists of the following fields:

- **titleId** (String): Unique alphanumeric identifier for a title.
- **ordering** (Integer): Specifies the sequence order of alternate titles.
- **title** (String): The alternate title for the movie or TV show.
- **region** (String): Two-letter region code representing the country or region.
- **language** (String): The language in which the title is expressed.
- **types** (String): Describes the type of alternate title (e.g., festival, DVD release, alternative title).
- **attributes** (String): Additional notes about the title variation (e.g., reissue title, working title).
- **isOriginalTitle** (Boolean): Indicates if the title is the original title (1) or an alternate (0).

Data Profiling Summary

The data profiling was conducted using Alteryx, and the key insights are as follows:

Field-Level Summary

Field Name	Data Type	Unique Values	Null %	Min Value	Max Value	Min Length	Max Length
titleId	**String**	**11,497,439**	**0%**	**tt0000001**	**tt10001000**	**9**	**10**

ordering	**String**	**251**	**0%**	**1**	**99**	**1**	**3**
title	**String**	**7,349,672**	**0%**	**B**	**Longest film title in dataset**	**1**	**254**
region	**String**	**249**	**0%**	**AD**	**CSHH**	**2**	**4**
language	**String**	**109**	**0%**	**\N**	**cmn**	**2**	**3**
types	**String**	**24**	**0%**	**\N**	**working**	**2**	**20**
attributes	**String**	**185**	**0%**	**16mm release title**	**weekend title**	**2**	**62**
isOriginalTitle	**String**	**2**	**0%**	**0**	**1**	**1**	**1**

Data Quality Assessment

Issue	**Details (Including Percent Value)**	**Alteryx Action**
Null Values	**Some fields contain \N (e.g., language, region, types).**	**Replace with NULL values for consistency.**
Data Type Issues	**ordering, isOriginalTitle stored as strings instead of integers.**	**Convert ordering and isOriginalTitle to Integer data type.**
Text Standardization	**title contains leading/trailing whitespace.**	**Remove unnecessary whitespace.**

4. Data Cleaning & Transformation Plan

To ensure data quality and usability, the following steps will be implemented:

- Convert ordering and isOriginalTitle to Integer data type.
- Replace \N values with NULL to handle missing data properly.
- Standardize text formatting in title.
- Validate the range of values in region and language fields.
- Ensure proper encoding of non-English characters in title.

5. Data Transformation & Standardization

Transformation Type	Description
Handling Missing Data	Replace \N values with NULL where applicable.
Data Type Conversion	Convert ordering and isOriginalTitle to integers.
Fixing Data Ranges	Validate region and language field values for consistency.
Standardizing Text Fields	Remove leading/trailing whitespace from title.
Removing Duplicates	Identify and eliminate duplicate records if any exist.

title.crew.tsv:

The *title.crew.tsv.gz* dataset is part of the IMDb Non-Commercial Datasets and provides information about the directors and writers associated with various titles in the IMDb database. This dataset is crucial for analyzing creative contributions to movies, TV series, and other types of content. By linking titles with their respective directors and writers, this dataset enables insights into creative patterns, collaborations, and trends in filmmaking.

This dataset will be used in this project to:

- Identify key contributors (directors and writers) across different genres and formats.
- Analyze collaboration networks between directors and writers.
- Integrate with other datasets (e.g., ratings, title details) to generate insights into correlations between crew involvement and audience reception.

2. Dataset Description

The dataset consists of the following fields:

- tconst (String): Unique alphanumeric identifier for a title.
- directors (Array of nconsts): Alphanumeric identifiers for the director(s) of the title.
- writers (Array of nconsts): Alphanumeric identifiers for the writer(s) of the title.

3. Data Profiling Summary

The data profiling was conducted using Alteryx, and key insights are summarized below:

Field-Level Summary

Field Name	Data Type	Unique Values	Null %	Min Value	Max Value	Min Length	Max Length
tconst	String	944829	0%	tt0031458	tt9916880	9	10

directors	String	1405422	0%	tt0041038	nm9993709	2	254
writers	String	11500529	0%	1	nm9993713,nm5411362,nm1744021	2	254

4.Data Quality Assessment

- Null Values: Records contain \N in the *directors* and *writers* fields to denote missing data and are replaced with NA
- Data Type Issues: All fields are stored as strings; no conversion issues detected.
- Low Frequency Values: Some fields have unique or rarely occurring values.

Text Standardization: No leading or trailing whitespace issues detected.

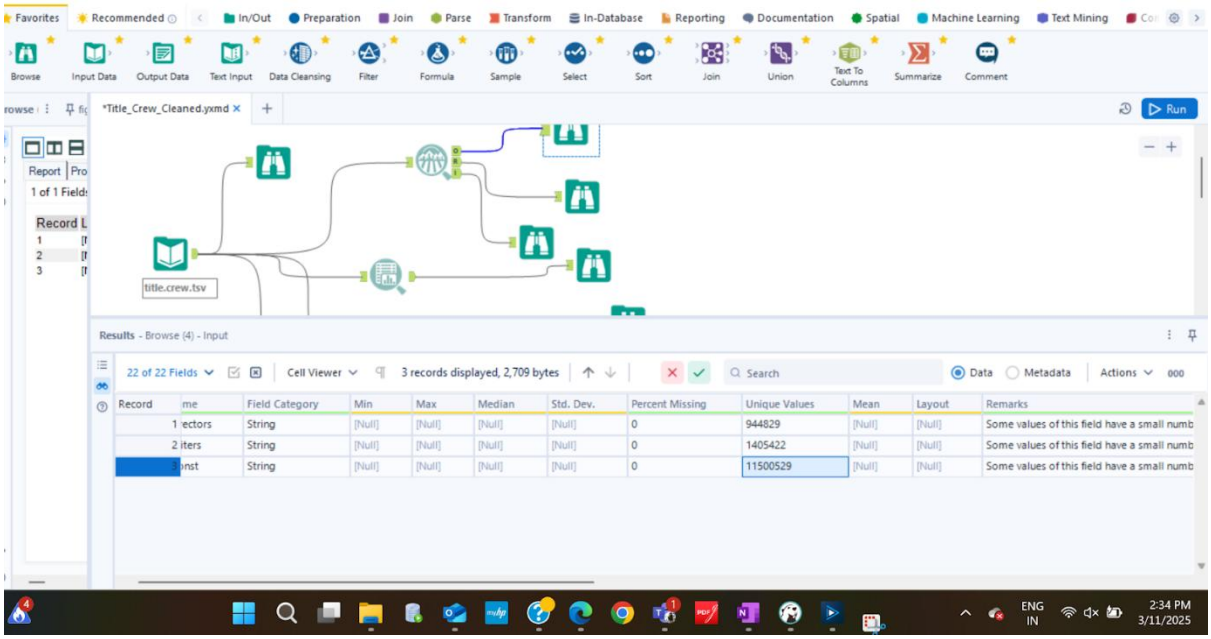
Data Cleaning & Transformation Plan

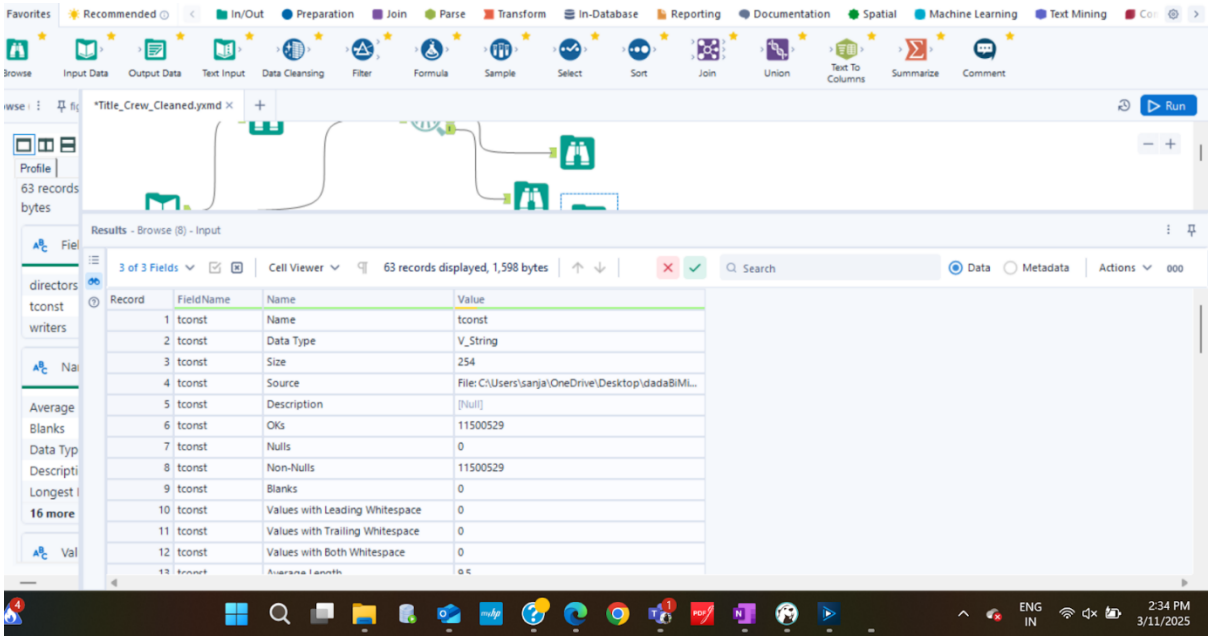
To ensure data quality and usability, the following steps will be implemented:

1. Replace \N values with NULL to handle missing data properly.
2. Validate lengths of *directors* and *writers* fields to ensure logical constraints (e.g., maximum number of contributors).
3. Group low-frequency values where appropriate to improve analysis.
4. Ensure consistent text formatting across all fields.

Transformation Type	**Description**
Handling Missing Data	Replace \N values with NULL where applicable.
Data Type Conversion	Ensure all fields are correctly formatted as strings for analysis purposes.

Standardizing Text Fields	Ensure consistency in text formatting and eliminate inconsistencies.
Removing Duplicates	Identify and eliminate duplicate records.





title.episodes.tsv:

Introduction

The title.episode.tsv.gz dataset is part of the IMDb Non-Commercial Datasets and provides episode-level details for TV series. It links individual episodes to their respective parent series, allowing for detailed analysis of season structures, episode counts, and TV show trends over time.

This dataset plays a crucial role in understanding television content by enabling business intelligence (BI) insights such as seasonal trends, episode distributions, audience preferences, and crew involvement across different TV series. The data is structured with unique identifiers for each episode and its corresponding parent TV series, along with season and episode numbers.

The dataset will be used in this project to:

- Track the number of episodes per season and series.
- Analyze the longevity and consistency of TV series.
- Integrate with other datasets (e.g., ratings, crew, and title details) to generate meaningful insights.

2. Dataset Description

The dataset consists of the following fields:

- tconst (String): Unique alphanumeric identifier for an episode.
- parentTconst (String): Unique alphanumeric identifier for the parent TV series.
- seasonNumber (Integer): The season number to which the episode belongs.
- episodeNumber (Integer): The episode number within the respective season.

3. Data Profiling Summary

The data profiling was conducted using Alteryx, and the key insights are as follows:

Field-Level Summary

Field Name	Data Type	Unique Values	Null %	Min Value	Max Value	Min Length	Max Length
tconst	String	8,844,677	0%	tt0031458	tt10001006	9	10
parentTconst	String	217,152	0%	tt0041038	tt32857063	9	10
seasonNumber	String	320	0%	1	1994	1	4
episodeNumber	String	15,785	0%	1	10312	1	5

Data Quality Assessment

Issue	Details (Including Percent Value)	Alteryx Action
Null Values	Some records contain \N in seasonNumber and episodeNumber.	Replace with NULL values to ensure consistency.
Data Type Issues	seasonNumber and episodeNumber are stored as strings.	Convert to Integer data type.
Low Frequency Values	Some fields have unique or rarely occurring values.	Consider grouping low-frequency values where appropriate.
Text Standardization	No leading or trailing whitespace issues detected.	Maintain consistent text formatting.

4. Data Cleaning & Transformation Plan

To ensure data quality and usability, the following steps will be implemented:

- Convert seasonNumber and episodeNumber to integer data types.
- Replace \N values with NULL to handle missing data properly.
- Ensure seasonNumber values are within a reasonable range (1 to a realistic maximum).
- Validate episodeNumber values against logical constraints.
- Consider grouping values with very low occurrences to improve analysis.

5. Data Transformation & Standardization

Transformation Type	Description
Handling Missing Data	Replace \N values with NULL where applicable.
Data Type Conversion	Convert seasonNumber and episodeNumber to integers.
Standardizing Text Fields	Ensure consistency in text formatting and eliminate inconsistencies.
Removing Duplicates	Identify and eliminate duplicate records.

title.ratings.tsv:

Introduction

The IMDb Ratings dataset provides user ratings and vote counts for a wide range of movies and TV shows. The dataset is crucial for understanding audience preferences, analyzing rating distributions, and conducting data-driven decision-making in the entertainment industry.

This report aims to **profile, assess, clean, and transform** the dataset to ensure high data quality and readiness for analysis.

2. Dataset Overview

The IMDb Ratings dataset contains **1,541,709** records and **3 columns**, providing details about **average user ratings and the number of votes received for various movie and TV show titles**.

Dataset Structure

Column Name	Data Type	Description
tconst	**String**	Unique IMDb identifier for a title
averageRating	**String**	IMDb average rating (1-10 scale)
numVotes	**String**	Number of votes the title received

3. Data Profiling & Quality Assessment

Missing Values Check

- **No missing values** were found in any column.
- **No \N placeholders** were detected.

Data Type Issues

- **averageRating & numVotes were initially stored as strings**, which is incorrect.
- **Solution:** Convert averageRating to **FLOAT**, and numVotes to **INTEGER**.

Duplicate Records Check

- tconst is **unique** across all records.
- **No duplicate records** found.

Statistical Summary & Key Insights

Column	Min	Max	Mean	Standard Deviation
averageRating	**1.0**	**10.0**	~6.9	Moderate variance
numVotes	**1**	**Millions**	**Few thousand**	Highly skewed

- **Right-skewed distribution**: Most movies have **low votes**, while a few titles receive **millions of votes**.
- **Outliers detected**: Some titles have very few votes, which may not be reliable.

4. Field-Level Summary

Each column is analyzed for its distribution, uniqueness, and validity.

tconst (Unique Title Identifier)

- **100% unique values**
- No null or duplicate values
- Used as the **primary key** for joining with other IMDb datasets.

averageRating (IMDb Rating: 1-10 Scale)

- **Range: 1.0 - 10.0**
- **Mean rating ~6.9**, indicating most movies are rated between **5.5 and 8.0**.
- **Outliers detected**: Some titles have **extreme ratings (1.0 or 10.0)**, which could indicate bias.

numVotes (Number of Votes per Title)

- **Highly skewed distribution**: Some movies have only **a few votes**, while blockbusters have **millions**.
- **Outliers detected**: Titles with fewer than **100 votes** may not be reliable.

5. Data Quality Assessment

Quality Check	**Assessment**	**Resolution**
Missing Values	No missing values	No action needed
Incorrect Data Types	averageRating & numVotes as strings	Convert to float & integer
Duplicate Records	No duplicates found	No action needed
Outliers in numVotes	High variance in votes	Consider filtering movies with < 100 votes

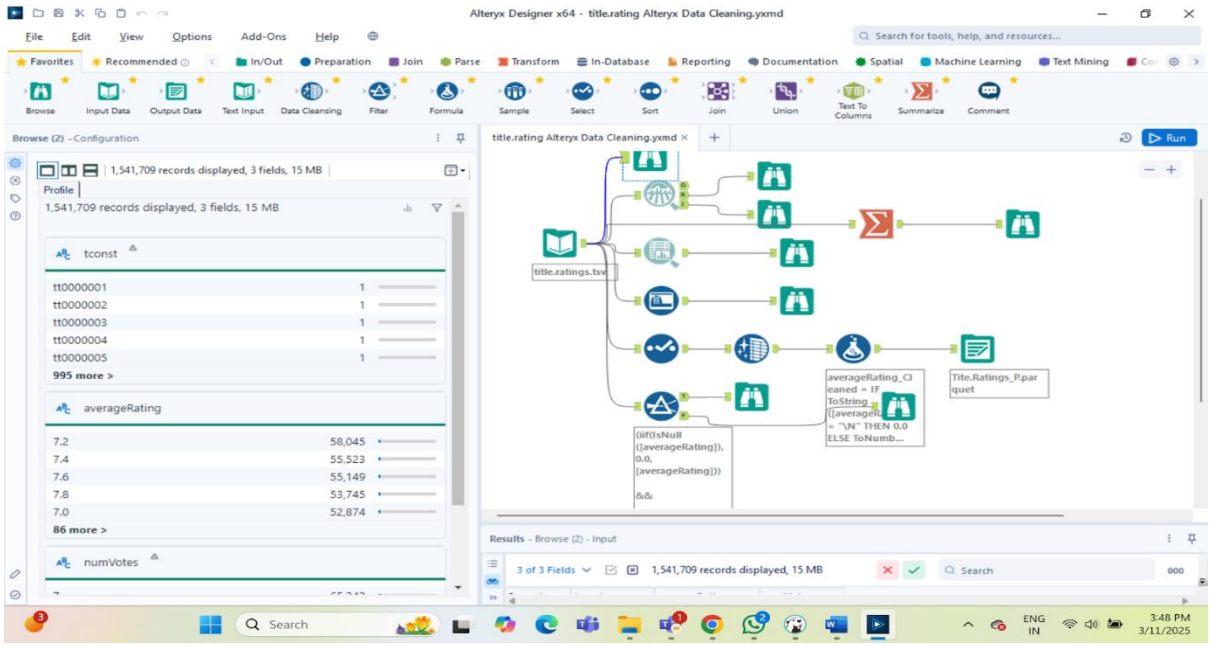
6. Data Cleaning & Transformation Plan

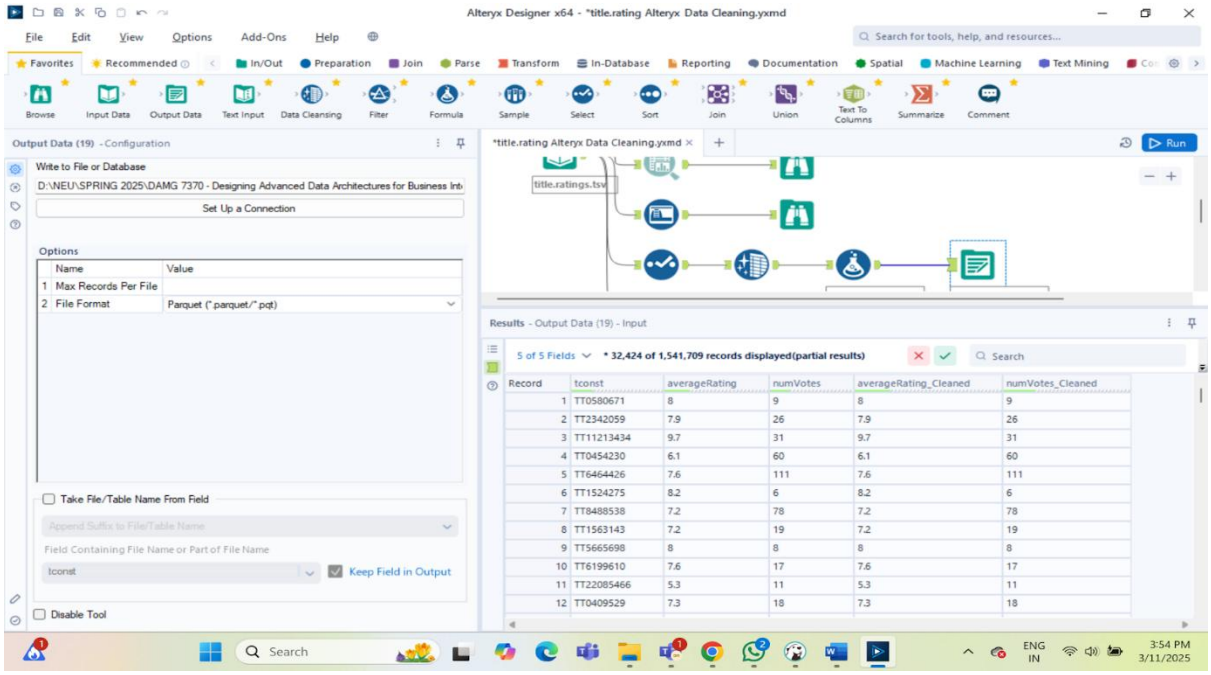
Convert Data Types

- Convert averageRating **String → FLOAT**.
- Convert numVotes **String → INTEGER**.

Handle Outliers

- **Low vote counts (numVotes < 10)** may indicate **unreliable ratings**.
- **Recommended Action:** Filter out titles with fewer than **100 votes**.


Alteryx Designer x64 - title.rating Alteryx Data Cleaning.yxmd
File Edit View Options Add-Ons Help
Search for tools, help, and resources...
Favorites Recommended In/Out Preparation Join Parse Transform In-Database Reporting Documentation Spatial Machine Learning Text Mining
Browse Input Data Output Data Text Input Data Cleansing Filter Formula Sample Select Sort Join Union Text To Columns Summarize Comment
Browse (2) - Configuration
1,541,709 records displayed, 3 fields, 15 MB
Profile
1,541,709 records displayed, 3 fields, 15 MB
tconst
tt0000001 1
tt0000002 1
tt0000003 1
tt0000004 1
tt0000005 1
995 more >
averageRating
7.2 58,045
7.4 55,523
7.6 55,149
7.8 53,745
7.0 52,874
86 more >
numVotes
title.rating Alteryx Data Cleaning.yxmd
Run
title.ratings.tsv
averageRating_Cleaned = IF ToString([averageRating]) = "\N" THEN 0.0 ELSE ToNumb...
Tite.Ratings_P.parquet
iif(IsNull([averageRating]), 0.0, [averageRating]))
&&
Results - Browse (2) - Input
3 of 3 Fields
1,541,709 records displayed, 15 MB
Search
ENG IN
3:48 PM 3/11/2025


Alteryx Designer x64 - *title.rating Alteryx Data Cleaning.yxmd
File Edit View Options Add-Ons Help
Search for tools, help, and resources...
Favorites Recommended In/Out Preparation Join Parse Transform In-Database Reporting Documentation Spatial Machine Learning Text Mining
Browse Input Data Output Data Text Input Data Cleansing Filter Formula Sample Select Sort Join Union Text To Columns Summarize Comment
Output Data (19) - Configuration
Write to File or Database
D:\NEU\SPRING 2025\DAMG 7370 - Designing Advanced Data Architectures for Business Int
Set Up a Connection
Options
Name Value
1 Max Records Per File
2 File Format Parquet (*.parquet;*.pqt)
Take File/Table Name From Field
Append Suffix to File/Table Name
Field Containing File Name or Part of File Name
tconst
Keep Field in Output
Disable Tool
*title.rating Alteryx Data Cleaning.yxmd
Run
title.ratings.tsv
Results - Output Data (19) - Input
5 of 5 Fields
* 32,424 of 1,541,709 records displayed(partial results)
Search
Record tconst averageRating numVotes averageRating_Cleaned numVotes_Cleaned
1 TT0580671 8 9 8 9
2 TT2342059 7.9 26 7.9 26
3 TT11213434 9.7 31 9.7 31
4 TT0454230 6.1 60 6.1 60
5 TT6464426 7.6 111 7.6 111
6 TT1524275 8.2 6 8.2 6
7 TT8488538 7.2 78 7.2 78
8 TT1563143 7.2 19 7.2 19
9 TT5665698 8 8 8 8
10 TT6199610 7.6 17 7.6 17
11 TT22085466 5.3 11 5.3 11
12 TT0409529 7.3 18 7.3 18
ENG IN
3:54 PM 3/11/2025