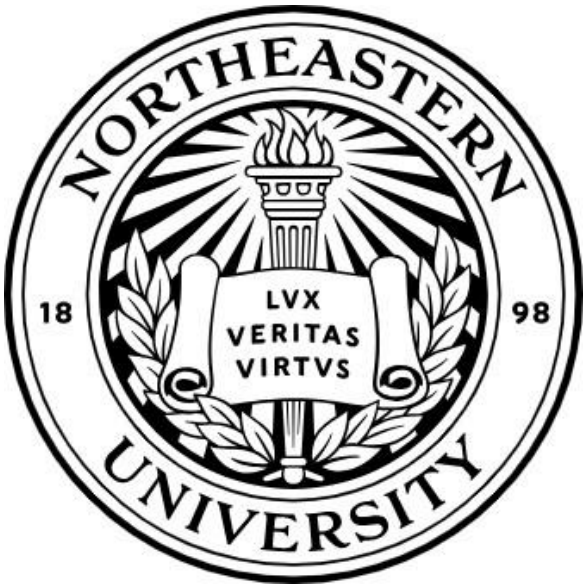
Northeastern University
College of Engineering



DAMG 7370: Designing Advanced Data Architectures for Business Intelligence

Prof: Naveen, Kuragayala

Final Project

Vikash Singh

OUR ASSUMPTIONS

1. All the dims in our model are of Type-1, including DIM_Title_Basics, but in order to demonstrate that we are aware of the Type-2 SCD component, we are implementing it as an orphan dimension.

2. Null Values are kept as -1 and are ignored when creating visualizations.

3. We have created a master list of all the people involved in the database as master_nconst_tconst_category table, and we have used it to create our DIM_Person, which gives us the Actor, Writer, and Director according to the Business Requirements.

4. The TSV source files are used for seasonal analysis in the visualizations

ALTERYX PROFILING

Sources

In this project we are using 3 different types of sources like MYSQL database, TSV files and then JSON files.
IMDB is a schema which contains all the tables related to movies, statistics, etc

IMDB Database schema:

imdb_tables_backup.sql

Tables Overview:

- Imdb_name_basics

1454602 records

Columns	DataType	Min	Min Max	Max	Max Values	Null	Description
nconst	String	9	nm0000001	10	nm10001020	Not null	Durable key used to identify each **person** in the data.
primaryName	String	1	Q	95	The Men of Battle Company 2nd of the 503rd Infantry Regiment 173rd Airborne Brigade Combat Team	Not null	Name of the person
birthYear	String	2	0004	4	\N	Not null	Year of Birth of individual
deathYear	String	2	0017	4	\N	Not null	Year of Death of individual
primaryProfession	String	5	actor	65	location_management, production_designer, transportation_department	Not null	Is a denormalized column which shows all the professions of the person.
knownForTitles	String	2	\N	59	tt2316835,tt2261689, tt2631704,tt2520738, tt2386718,tt2267526	Not null	Denormalized column which

- Imdb_title_akas

2,612,798 records

Columns	Description	Min	Min Value	Max	Missing Values	Null	Description
titleId	String	9	tt0000502	10	tt10001116	Not null	Durable key, tconst, used to identify movies
ordering	String	1	1	3	149	Not null	Used to check which movie was released first
title	String	1	B	273	En cuanto a la infancia, andar, regeneración y muerte de Jorge del Carmen Valenzuela Torres, quien se hace llamar también José del Carmen Valenzuela Torres, Jorge Sandoval Espinoza, José Jorge Castillo Torres, alias El Campano, El Trucha, El Canaca, El ...	Not null	Name of the movie
region	String	2	AD	4	CSHH	Not null	Release Region
language	String	2	\N	3	qbn	Not null	Language in which it was released
types	String	2	\N	20	festivalimdbDisplay	Not null	
attributes	String	2	\N	62	added framing sequences and narration in Yiddishreissue title	Not null	
isOriginalTitle	String	2	\N	2	\N	Not null	

- Imdb_title_basics

607,423 records

Columns	DataType	Min	Min Values	Max	Max Values	Null	Description
tconst	String	9		10	N/A	Not null	Durable Key
titleType	String	5		5	N/A	Not null	
primaryTitle	String	1	M	242	I Saw a Little Bird Flying Over a Psychiatric Hospital Near Milan 2 Days Ago and Now it Became So Good that I Drank the Bordeaux Champagne of the Sample in 1771 So That My Soul Would Experience an Incredible Life in This Brutal and Gray World	Not null	Title of the movie
originalTitle	String	1	M	242	I Saw a Little Bird Flying Over a Psychiatric Hospital Near Milan 2 Days Ago and Now it Became So Good that I Drank the Bordeaux Champagne of the Sample in 1771 So That My Soul Would Experience an Incredible Life in This Brutal and Gray World	Not null	Original Title of the movie
isAdult	Int	0	0	1	1	Not null	0 for False, 1 for True
startYear	String	2	1905	4	\N	Not null	Year of movie released
endYear	String	2	\N	2	\N	Not null	Year of movie elapse
runtimeMinutes	String	1	2	5	10062	Not null	Runtime of movies
genres	String	2	Action	32	Biography,Documentary,Reality-TV	Not null	Denormalized column consisting of the movie genres

- Imdb_title_crew

607,423 records

Columns	DataType	Min	Min Values	Max	Max Value	Null	Description
tconst	String	9	tt0000502	10	tt10001116		Durable Key referencing the imdb_title) basics table
directors	String	2	\N	861	nm7832585,nm6376917, nm7841280,nm9075411, nm3530863,nm8329010, nm1468573,nm9744786, nm9425998,nm3683211, nm6642750,nm3894671, nm5169423,nm8487362, nm3549368,nm8734809, nm4677186,nm2358966, nm7429538,nm8575753, nm8215796,nm8660970, nm8381975,nm5755088, nm8657022,nm8...		Denormalized column consisting of Nconst(Durable Key from the name basics table
writers	String	Not null	2	1189	nm0000601,nm0000040, nm0000101,nm0000118, nm0000175,nm0000233, nm0000339,nm0000341, nm0000484,nm0000600, nm0000783,nm0000812, nm0001053,nm0001054, nm0001094,nm0001140, nm0001252,nm0001279, nm0001361,nm0001469, nm0001548,nm0001873,		Denormalized column consisting of Nconst(Durable Key from the name basics table)

					nm0002657,nm0005196, nm0005351		

- Imdb_title_principals

4,283,620 records

Columns	DataType	Min	Min Values	Max	Max Value	Null	Description
tconst	String	9	tt0000502	10	tt10001116		Durable Key referencing the imdb_title) basics table
ordering	String	1	1	2	10		
nconst	String	9	nm0000001	10	nm11468708		Durable Key, referencing
category	String	4	self	19	production_designer		Denormalized Column containing the category
job	String	1	h	157	works "Visions of Gerard", "Dr. Sax", "Vanity of Duluoz", "The Town and the City", "On the Road", "Desolation Angels", "Dharma Bums" and "Big Sur and others"		-NA
characters	String	2	\N	177	["(segments \"Chanel No.5: The Swimming Pool\" - \"Chanel No.5: Monuments\" - \"Chanel No. 5: La Star\" - \"Chanel No. 5: Sentiment troublant\" - \"Chanel No.5: L'orchestre\")"]		-NA

- Imdb_title_ratings

277,171 records

Columns	Description	Min	Min Values	Max	Max Values	Null	Description
tconst	String	9	tt10002654	10	tt9916730	Not null	Durable Key referencing the imdb_title) basics table
averageRating	Double	1	1	3	10	Not null	Average Rating of the movies
numVotes	Int	1	5	7	2572812	Not null	Votes received by the movie

Numbers.tsv: In this folder there are 9 TSV files which are loaded into single Numbers_tsv table In MySQL

TheNumbers - Moviewise Daily BoxOffice Revenues
2930 records

Columns	DataType	Min	Min Values	Max	Max Value	Null	Description
tconst	String	9	tt0120338	10	tt10872600		Durable Key referencing the imdb_title) basics table
title	String	6	Avatar	42	Star Wars: Episode VII - The Force Awakens		Name of the Movie
Date	Date	10	2022-04-22	10	1997-12-19		Released Date
Rank	String	1	1	2	10		-NA
Gross	String	4	$571	12	$157,461,641		Daily Gross Collection of the movie
_YD	String	2	2	4	-34%		-NA
_LW	String	3	-2%	4	-14%		-NA
Theaters	String	1	9	5	3,452		Number of theatres its released
Per_Theater	String	3	$60	7	$19,744		Revenue per theatre
Total_Gross	String	10	$8,658,814	12	$109,497,762		Total gross for each day for each movie
Days	String	1	1	5	5,221		-NA

Json: In this we have 2 JSON files with new title and names tables data.

SCD2 Data

new_name_basics.json:

7 records

Columns	DataType	Min	Min Value	Max	Max Value	Null	Description
nconst	String	9	nm0000375	9	nm4043618	Not null	N/A
primaryName	String	23	Mark Ruffalo (aka Hulk)	38	Benedict Cumberbatch (aka Dr. Strange)	Not null	N/A
birthYear	String	4	1995	4	1996	Not null	N/A
deathYear	String	2	\N	2	\N	No Values	No Values
primaryProfession	String	23	actor,director,producer	27	actress,soundtrack, producer	Not null	N/A
knownForTitles	String	39	tt0371746,tt4154796, tt1300854,tt0988045	39	tt2395427,tt0458339, tt3498820,tt0848228	Not null	N/A

new_title_basics.json:

10 records

Columns	Data Type	Min	Min Value	Max	Max Value	Null	Description
tconst	String	9	tt0120338	10	tt10872600	Not null	Durable Key referencing the imdb_title) basics table
titleType	String	5	movie	5	movie	Not null	-NA
primaryTitle	String	13	Avatar (2009)	49	Star Wars: Episode VII - The Force Awakens (2015)	Not null	Title of the movie

originalTitle	String	6	Avatar	42	Star Wars: Episode VII - The Force Awakens	Not null	Original Title of the movie
isAdult	String	0	0	1	1	Not null	If its an A rated movie, its 1, otherwise it 0
startYear	String	4	1997	4	2021	Not null	Movie released year
endYear	String	2	\N	2	\N	Not null	-NA
runtimeMinutes	String	3	134	3	194	Not null	Runtime of the movie
genres	String	13	Drama,Romance	24	Action,Adventure,Fantasy	Not null	Genres of the movie

Observations:

Inconsistent Datatypes:

The dataset contains columns with incorrect data types. Ensuring that data types match the nature of the information they hold is crucial for accurate analysis and visualization.

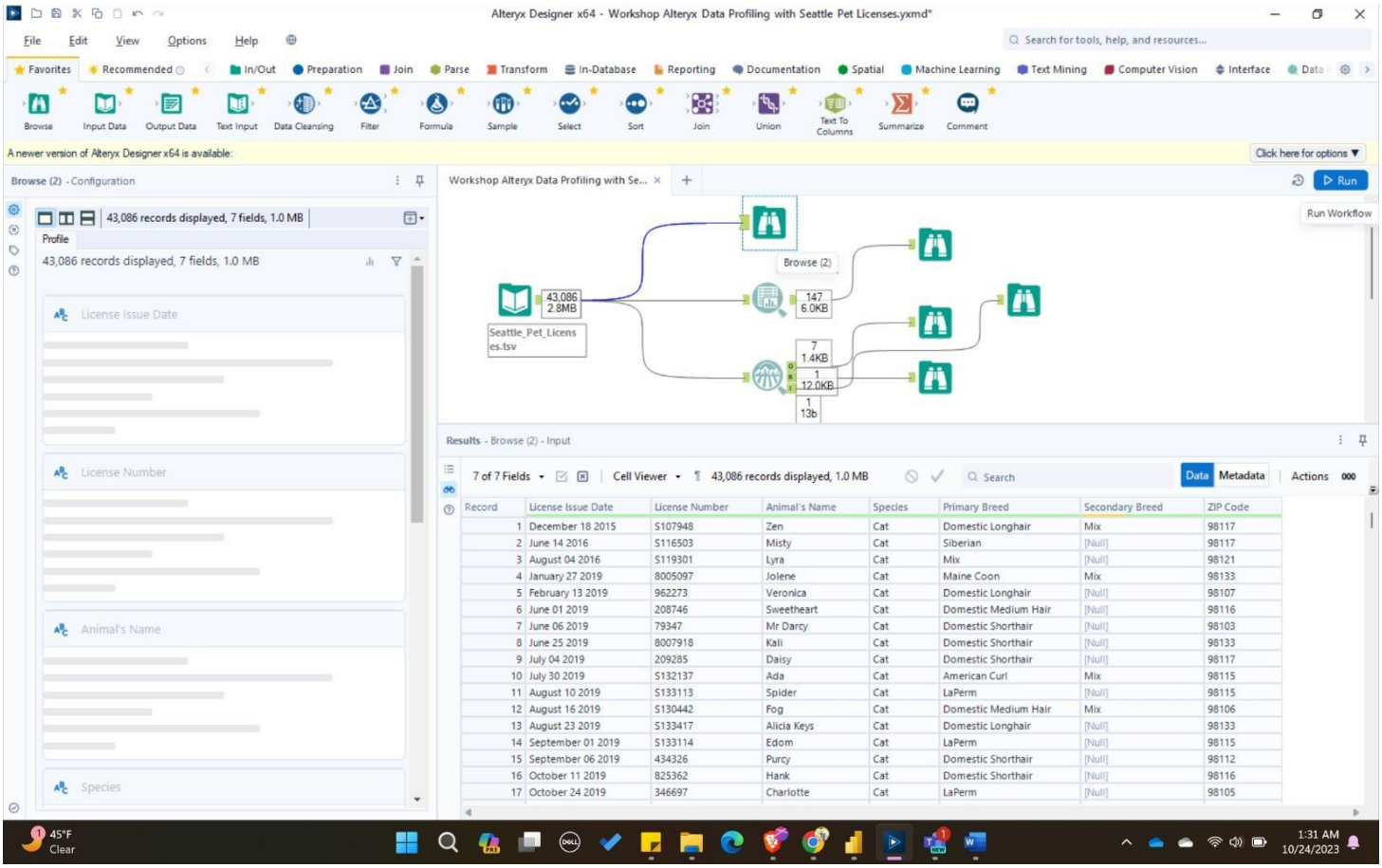
Multi-Valued records:

Three columns in the dataset contain more missing values. Identifying the reason for these gaps in data and addressing them effectively is critical to avoid skewed or incomplete analysis results.

Data Profiling:

We have Started by profiling the dataset to gain a clear understanding of the issues. Using Alteryx data profiling features to identify incorrect data types and null value percentages in each column.

Creating metadata in Talend to store information about the dataset's structure, including data types, null percentages, and any other relevant details.



Record	License Issue Date	License Number	Animal's Name	Species	Primary Breed	Secondary Breed	ZIP Code
1	December 18 2015	S107948	Zen	Cat	Domestic Longhair	Mix	98117
2	June 14 2016	S116503	Misty	Cat	Siberian	[Null]	98117
3	August 04 2016	S119301	Lyra	Cat	Mix	[Null]	98121
4	January 27 2019	8005097	Jolene	Cat	Maine Coon	Mix	98133
5	February 13 2019	962273	Veronica	Cat	Domestic Longhair	[Null]	98107
6	June 01 2019	208746	Sweetheart	Cat	Domestic Medium Hair	[Null]	98116
7	June 06 2019	79347	Mr Darcy	Cat	Domestic Shorthair	[Null]	98103
8	June 25 2019	8007918	Kali	Cat	Domestic Shorthair	[Null]	98133
9	July 04 2019	209285	Daisy	Cat	Domestic Shorthair	[Null]	98117
10	July 30 2019	S132137	Ada	Cat	American Curl	Mix	98115
11	August 10 2019	S133113	Spider	Cat	LaPerm	[Null]	98115
12	August 16 2019	S130442	Fog	Cat	Domestic Medium Hair	Mix	98106
13	August 23 2019	S133417	Alicia Keys	Cat	Domestic Longhair	[Null]	98133
14	September 01 2019	S133114	Edom	Cat	LaPerm	[Null]	98115
15	September 06 2019	434326	Purcy	Cat	Domestic Shorthair	[Null]	98112
16	October 11 2019	825362	Hank	Cat	Domestic Shorthair	[Null]	98116
17	October 24 2019	346697	Charlotte	Cat	LaPerm	[Null]	98105

Handling Incorrect Data Types:

For columns with incorrect data types, we used Talend's data transformation capabilities to cast or convert the data to the correct types.

Create data type conversion routines or functions in my Talend job to ensure data consistency.

Standardizing Unstructured Data:

For columns with unstructured formats (e.g., dates and times), will use Talend's data transformation functions to format the data consistently. For example, I can convert date and time strings to a standardized format.

Create Talend routines or expressions to parse and reformat these columns.

DIMENSIONAL MODEL