



MOTIVATION:

Emotions help us feel human again and connect us to each other by watching the lives of different characters and feel everything they feel through 24 frames per seconds of the movies. In this way, movies stir up your emotions and is one of the great things about them.

I personally a movie lover and sometimes I call myself as a cinephile who is passionate about movies and want to know a lot about them.

Moreover, a cinephile should be an educated film consumer with the tool kit to distinguish average films from outstanding ones depending on some particular metrics.

With that motivation, I create my data engineering capstone which we can quickly see the top 36 movie rental from [Cineplex website](#) along with the imdb rating, tomatometer and audience score from [Rotten tomatoes](#). In addition, dedicating to the “nerdy” movie lover, we can look up the cast in detail with the curated data sets from imdb.

IDEA OF THE PROJECT


When we click into the Top Rentals in Cineplex website, we might be confused and indecisive before choosing the right movie that you like.

This is when the project Top Rentals Cineplex comes in handy, it can quickly show you the essential information about the movie you want to rent:


- IMDB rating
- Tomato meter
- Audience Score
- Synopsis
- Top critics (from Rotten Tomato)

For my more demo visualizations please click this link below:

toprentalcineplex.my.canva.site

 **TOP RENTALS CINEPLEX**


Spider-Man: No Way Home

 **TOP 14**

IMDB RATING:	8.3
TOMATO METER:	93%
AUDIENCE SCORE:	98%

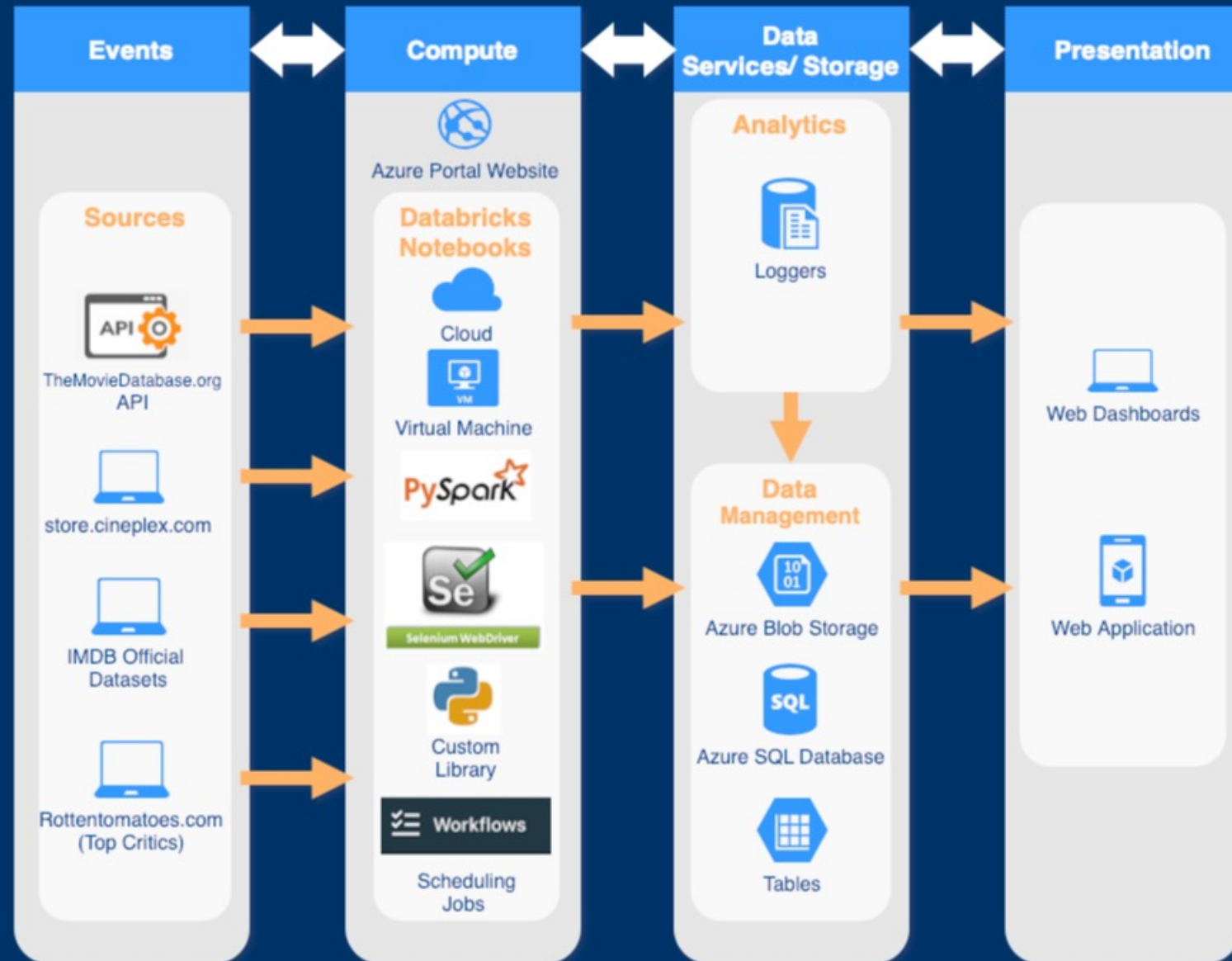
SYNOPSIS

TOP CRITICS



PROJECT
ETL
DIAGRAM

TOP RENTAL CINEPLEX PROJECT WITH MICROSOFT AZURE CLOUD SERVICES



EVENTS



The project data is collected/scraped from multiple sources by using Selenium WebDriver and API :

1. [Top Rentals Cineplex](#) : is the source to be scraped Top 36 Movie Rentals' title on Cineplex website.
2. [IMDB.com](#) : Official subsets of IMDB data that are available for personal and commercial use. IMDB data sets contain "imdb_id" which is used as primary/foreign key to link tables.
Following downloaded list:
 - **title.basics.tsv.gz** contains the essential information for the movie titles.
 - **title.crew.tsv.gz** contains the director and writer information for all the titles.
 - **title.principals.tsv.gz** contains the principal cast/crew for titles
 - **title.ratings.tsv.gz** contains the IMDB rating and votes information for titles
 - **name.basics.tsv.gz** contains the following information for names (actors, actresses, directors, writers, etc..)
3. [Themoviedb.org](#) : is a community built movie and TV database which has API available for everyone to use. I personally use their API to cumulate the movies' synopsis along with the "imdb_id" for the Top Movie Rentals from Cineplex.
4. [Rottentomatoes.com](#) : is used to achieve data for corresponding Top 36 Movie Rentals' title :
 - The top critics from credible reviewers or audience's review (for titles that don't have many credible reviewers).
 - Tomatometer and audience score.

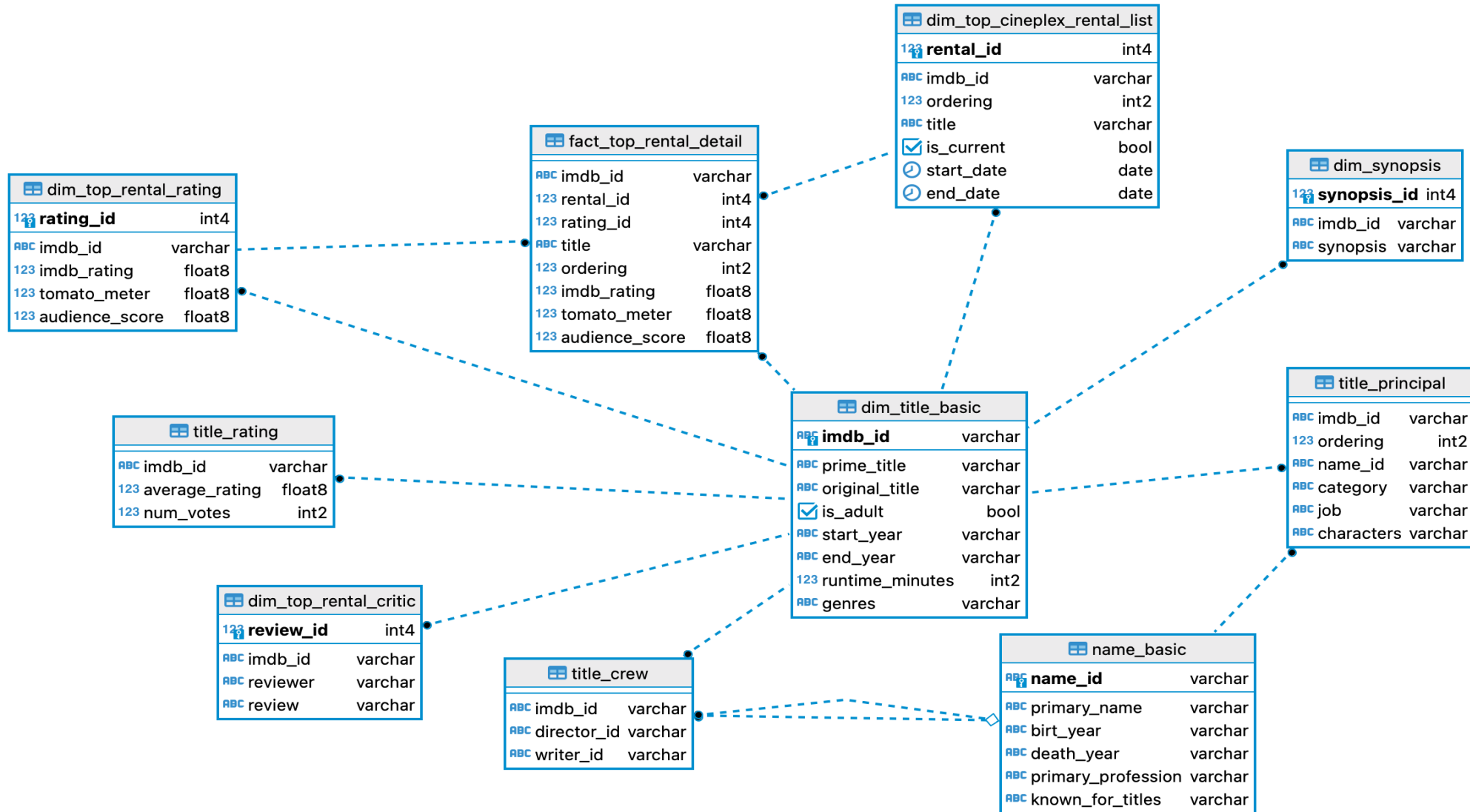
CLOUD COMPUTING SOLUTION

- I use Azure Databricks which is a fast, easy and collaborative Apache Spark-based big data analytics service designed for data science and data engineering.
- I created 4 notebooks in Databricks and incorporated some customized libraries:
 - **imdb_datasets_downloader** :
 - Automates the official data sets downloading process from IMDB website. (using Selenium)
 - Extracts **gz** files and convert **tsv** files to **parquet** files. (using PySpark)
 - Saves files to Azure Blob Storage (**ABS**) as data warehouse.
 - **top_rentals_cineplex_scrapper** :
 - Scrapes Top 36 movie rentals' titles on Cineplex website save as parquet file to **ABS** (using PySpark).
 - Applies Slowly Changing Dimension Type 2 for table structure that stores and manages the current and historical data over time in terms of the top titles orders (e.g: Top 1, 2 ,3 ,.. and the data is current or not current with date, time)
 - **theMovieDb_and_RottenTomatoes_data_scraper** :
 - Scrapes synopsis from themoviedb.org (using API) and top critics from rottentomatoes.com (using Selenium).
 - Saves table as parquet file to **ABS** (using PySpark).

CLOUD COMPUTING SOLUTION (CONTINUED)

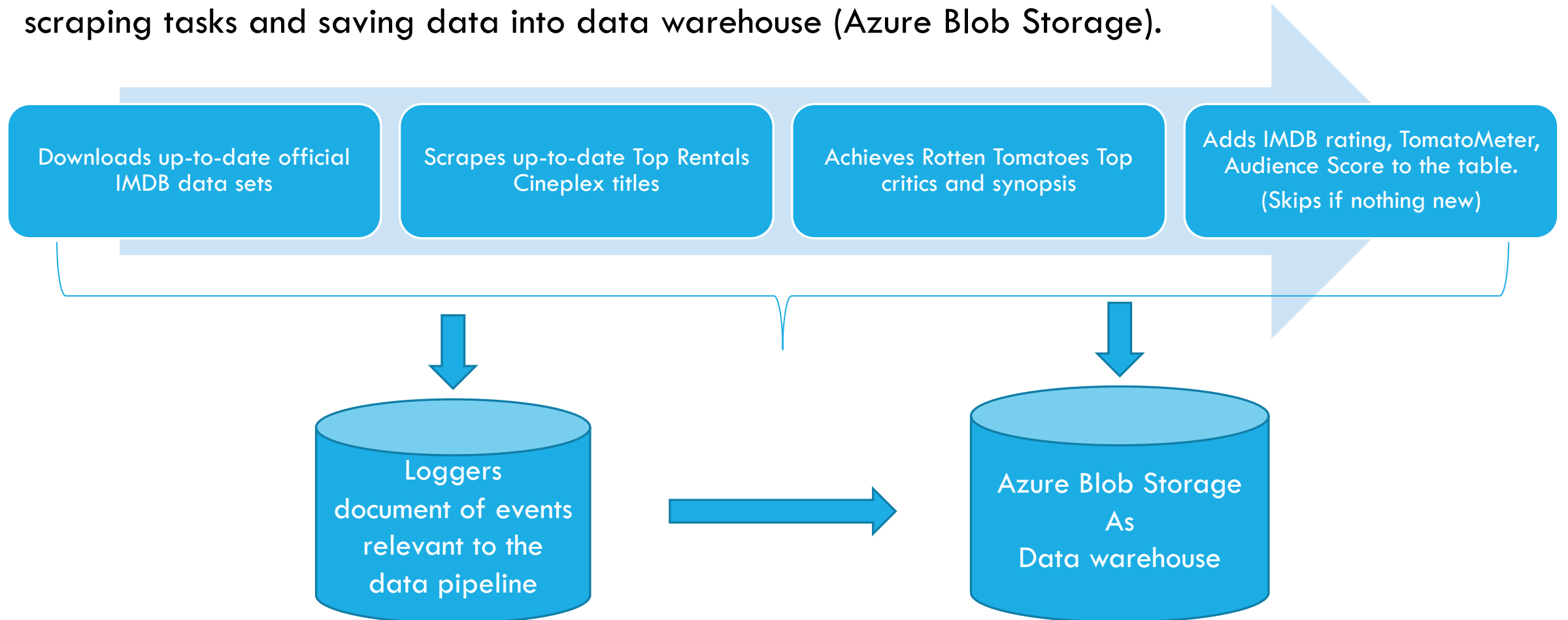
- **top_rental_rating_from_imdb_and_rotten_tomatoes_data :**
 - Extracts **imdb_rating** from IMDB data set and merge with Tomatometer and Audience Score from rottentomatoes.com into 1 table with corresponding **imdb_id** of the Top 36 movie rentals (using PySpark).
 - Saves as parquet file to **ABS**.

TOP RENTALS CINEPLEX'S ER DIAGRAM



DATA MANAGEMENT

- Azure Databricks Workflows is used to manage the scheduling jobs which automate the data scraping tasks and saving data into data warehouse (Azure Blob Storage).



DEMO DATA OUTPUT

Using PySpark to run the query for the fact table:

```
top_rental_cineplex.join(top_rental_rating, (top_rental_cineplex.imdb_id == top_rental_rating.imdb_id) , how = 'inner' ).\
    join(synopsis_table, (synopsis_table.imdb_id == top_rental_cineplex.imdb_id), how = 'inner').\
    filter((top_rental_cineplex.is_current == 1) ).\
    select(top_rental_cineplex.title,\
           top_rental_cineplex.ordering,\
           synopsis_table.synopsis,\
           top_rental_rating.imdb_rating ,\
           top_rental_rating.tomato_meter ,\
           top_rental_rating.audience_score

    ).orderBy(top_rental_cineplex.ordering).show()
```

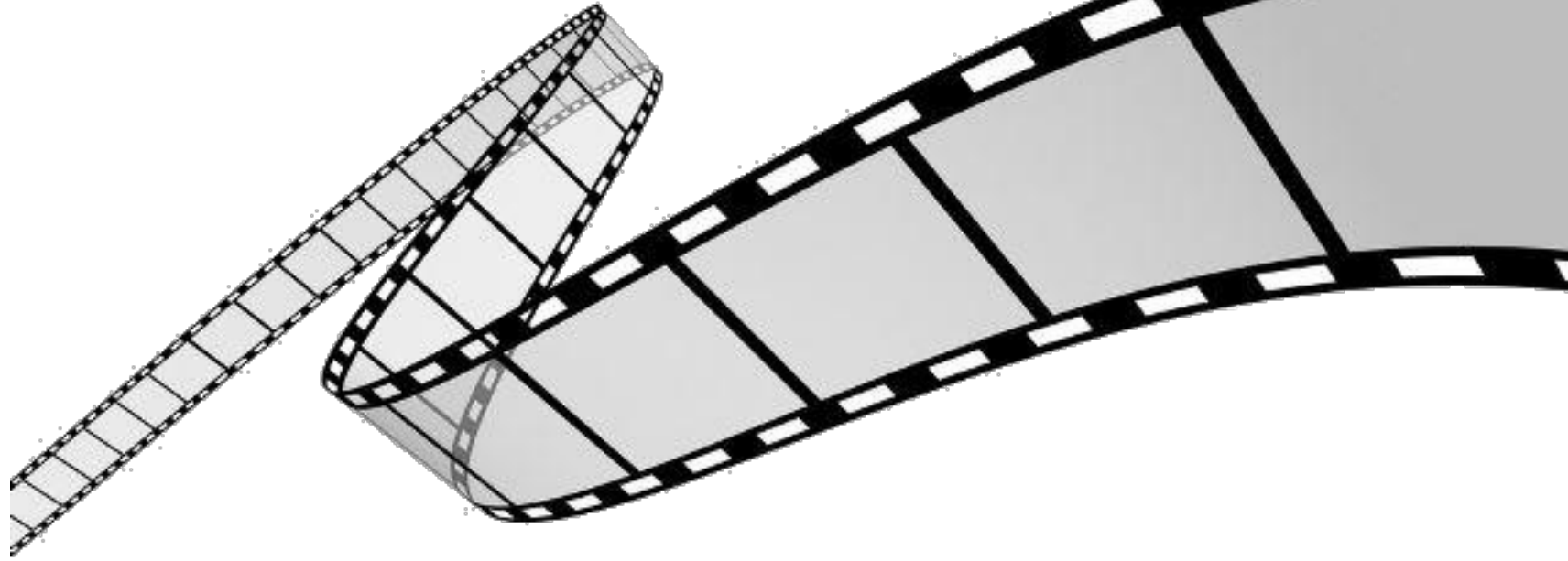
DEMO DATA OUTPUT (CONTINUED)

title	ordering	synopsis	imdb_rating	tomato_meter	audience_score
THE LOST CITY	1	A reclusive roman...	6.6	79	83
UNCHARTED	2	A young street-sm...	6.6	41	90
UMMA	3	Amanda and her da...	4.7	29	51
DOG	4	An army ranger an...	6.5	76	89
9 BULLETS	5	A former burlesqu...	null	0	41
STUDIO 666	6	Legendary rock ba...	5.8	56	80
MARRY ME	7	Music superstars ...	6.1	60	92
DC SHOWCASE: CONS...	8	John Constantine ...	null	null	29
SPIDER-MAN: ALL R...	9	Join our hosts JB...	null	null	null
JACKASS FOREVER	10	Celebrate the joy...	7.0	86	91
SING 2	11	Buster and his ne...	7.5	71	98
GHOSTBUSTERS: AFT...	12	When a single mom...	7.2	63	94
THE CURSED	13	In the late 19th ...	6.3	16	30
SPIDER-MAN: NO WA...	14	Peter Parker is u...	8.4	93	98
NITRAM	15	Based on true eve...	7.2	91	84
DEATH ON THE NILE	16	Belgian sleuth He...	6.3	62	82
HOUSE OF GUCCI	17	When Patrizia Reg...	6.6	62	83
NO TIME TO DIE	18	Bond has left act...	7.3	83	88

FUTURE WORK

- Creates Web Application to showcase the data.
- Completes Azure SQL database structure.
- Optimizes and refactor code for consistency and efficiency.

THANK YOU!



linkedin.com/in/andyphamto/



github.com/andy-pham-72/