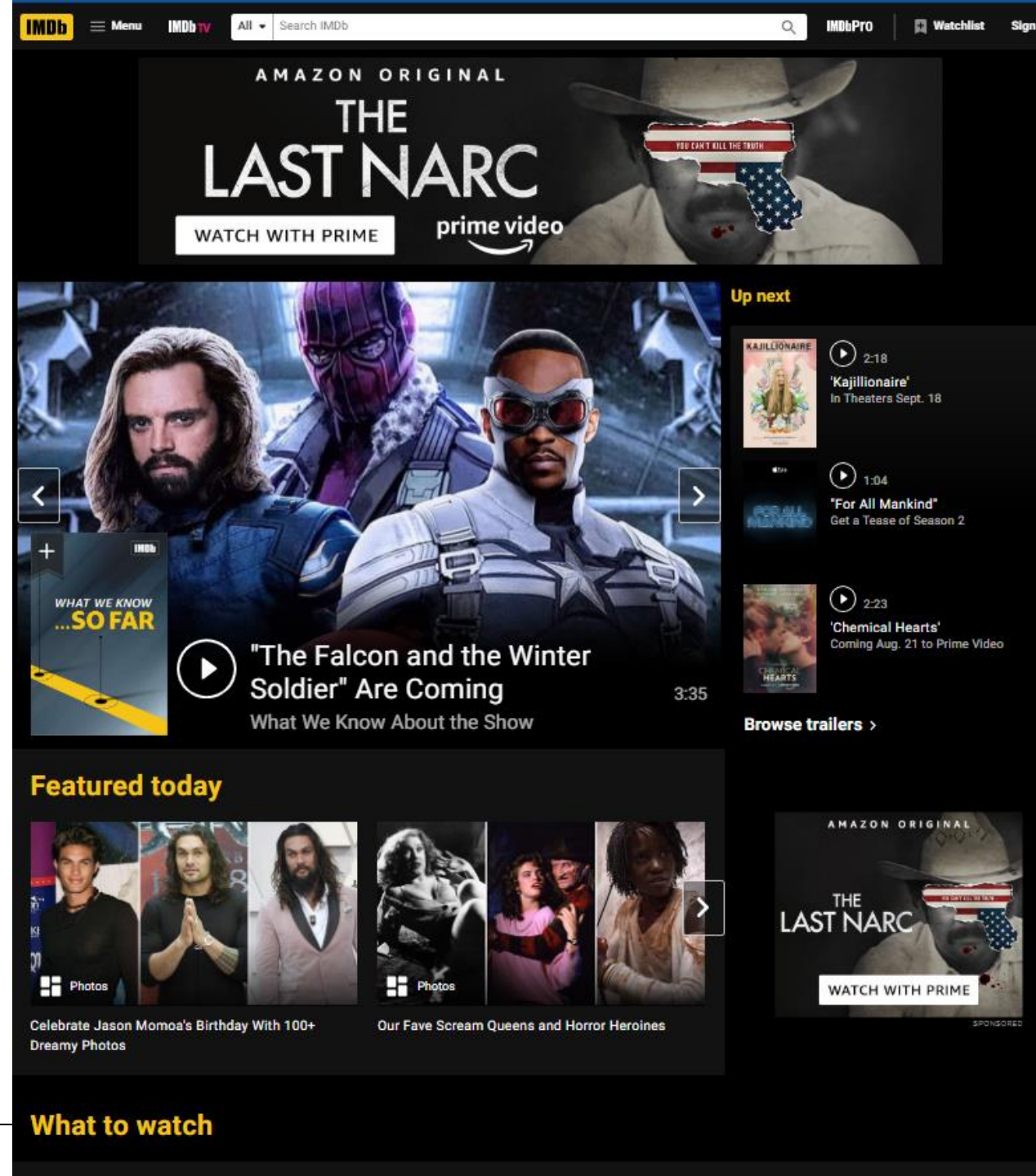


**IMDb**

# IMDb Data & Analysis Project



**Rick Sherman**  
Athena IT Solutions  
rick.sherman@athena-solutions.com



# IMDb

## What is IMDb?

“IMDb is the world's most popular and authoritative source for movie, TV and celebrity content, designed to help fans explore the world of movies and shows and decide what to watch.”

“Our searchable database includes millions of movies, TV and entertainment programs and cast and crew members. We help you jog your memory about a movie, show, or person on the tip of your tongue, find the best movie or show to watch next, and empower you to share your entertainment knowledge and opinions with the world’s largest community of fans.”

IMDb launched online in 1990 and has been a subsidiary of Amazon.com since 1998.

# IMDb Project

IMDb has an extensive database of movies and TV shows with associated actors, directors, writers and crew that is free to download. Although movies are on a bit of a hiatus this data offers interesting data integration and analytics challenges.

In addition, may potentially supplement the above datasets with:

- Various lists “cut & paste” from IMDb such as top box office (by revenue), top rated movies & tv shows, franchises, brands, awards, and other lists
- Box office, brands, franchises, genres and other lists from IMDb Pro [Box Office Mojo](#)
- Box office, franchises, distributors and other lists from [The Numbers](#)

# IMDb Project

A great feature of these datasets is that you can click on the title or name link and get to an IMDb page to examine the data and validate your results.

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE | SHARE

## + Resident Evil (2002)

R | 1h 40min | Action, Horror, Sci-Fi | 15 March 2002 (USA)

★ 6.7 / 10  
242,644

☆ Rate This



2:15 | Trailer | 2 VIDEOS | 165 IMAGES

A special military unit fights a powerful, out-of-control supercomputer and hundreds of scientists who have mutated into flesh-eating creatures after a laboratory accident.

**Director:** Paul W.S. Anderson  
**Writer:** Paul W.S. Anderson  
**Stars:** Milla Jovovich, Michelle Rodriguez, Ryan McCluskey | [See full cast & crew »](#)

[Watch on Prime Video](#) rent/buy from \$2.99

... [+ Add to Watchlist](#)

**33** Metascore From metacritic.com | **Reviews** 1,178 user | 238 critic | **Popularity** 1,363 (▲ 36)

### Cast

Cast overview, first billed only:

	<a href="#">Ryan McCluskey</a>	...	<a href="#">Mr. Grey</a>
	<a href="#">Oscar Pearce</a>	...	<a href="#">Mr. Red</a>
	<a href="#">Indra Ové</a>	...	<a href="#">Ms. Black</a>
	<a href="#">Anna Bolt</a>	...	<a href="#">Dr. Green</a>
	<a href="#">Joseph May</a>	...	<a href="#">Dr. Blue</a>
	<a href="#">Robert Tannion</a>	...	<a href="#">Dr. Brown</a>
	<a href="#">Heike Makatsch</a>	...	<a href="#">Lisa</a>
	<a href="#">Jaymes Butler</a>	...	<a href="#">Clarence</a>
	<a href="#">Stephen Billington</a>	...	<a href="#">Mr. White</a>
	<a href="#">Fiona Glascott</a>	...	<a href="#">Ms. Gold</a>
	<a href="#">Milla Jovovich</a>	...	<a href="#">Alice</a>
	<a href="#">Eric Mabius</a>	...	<a href="#">Matt</a>
	<a href="#">Colin Salmon</a>	...	<a href="#">One</a>
	<a href="#">Martin Crewes</a>	...	<a href="#">Kaplan</a>
	<a href="#">Pasquale Aleardi</a>	...	<a href="#">J.D.</a>

### Details

**Official Sites:** [Official Facebook](#) | [Sony Pictures \[United States\]](#)  
**Country:** [UK](#) | [Germany](#)  
**Language:** [English](#)  
**Release Date:** 15 March 2002 (USA) [See more »](#)  
**Also Known As:** [Resident Evil the Movie](#) [See more »](#)  
**Filming Locations:** [Kramnitz, Brandenburg, Germany](#) [See more »](#)

### Box Office

**Budget:** \$33,000,000 (estimated)  
**Opening Weekend USA:** \$17,707,106, 17 March 2002  
**Gross USA:** \$40,119,709  
**Cumulative Worldwide Gross:** \$102,984,862  
[See more on IMDbPro »](#)

### Company Credits

**Production Co:** [Constantin Film](#), [New Legacy](#), [Davis-Films](#) [See more »](#)  
[Show more on IMDbPro »](#)

### Technical Specs

**Runtime:** 100 min  
**Sound Mix:** [DTS](#) | [Dolby Digital](#) | [SDDS](#)  
**Color:** [Color](#)  
**Aspect Ratio:** 1.85 : 1  
[See full technical specs »](#)



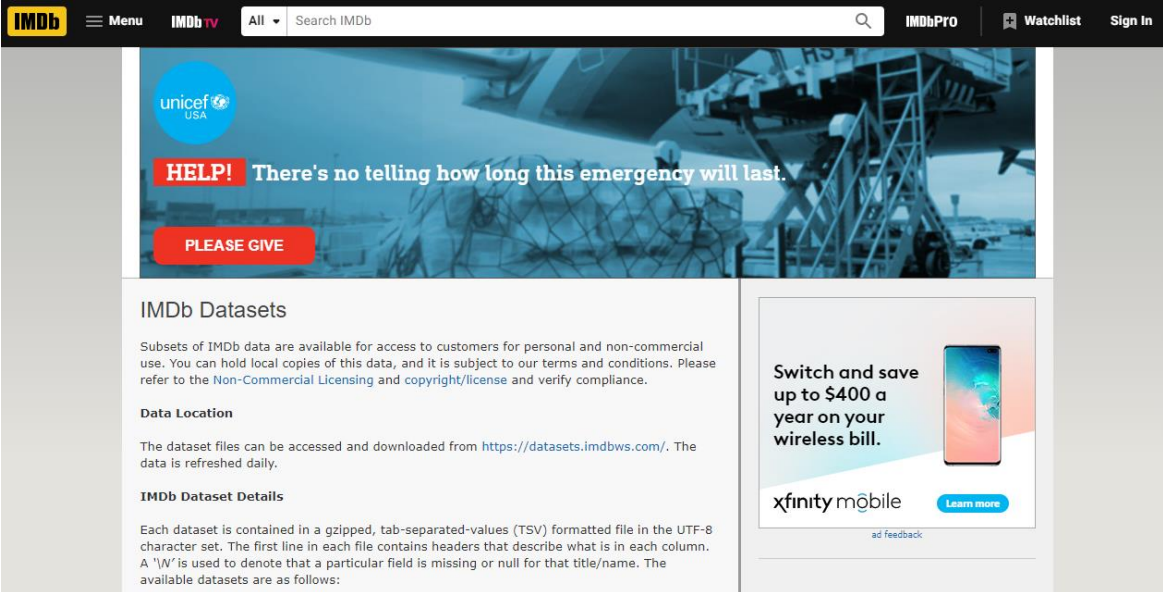
# IMDb Datasets

## IMDb Datasets

Subsets of IMDb data are available for access to customers for personal and non-commercial use.

## Data Location

The dataset files can be accessed and downloaded from <https://datasets.imdbws.com/>. The data is refreshed daily.



The screenshot shows the IMDb website's 'IMDb Datasets' page. At the top, there's a navigation bar with the IMDb logo, a menu icon, 'IMDb TV', a search bar, and links for 'IMDbPro', 'Watchlist', and 'Sign In'. Below the navigation bar is a large banner for UNICEF USA with the text 'HELP! There's no telling how long this emergency will last.' and a 'PLEASE GIVE' button. The main content area is titled 'IMDb Datasets' and contains the following text: 'Subsets of IMDb data are available for access to customers for personal and non-commercial use. You can hold local copies of this data, and it is subject to our terms and conditions. Please refer to the Non-Commercial Licensing and copyright/license and verify compliance.' Below this, under the heading 'Data Location', it states: 'The dataset files can be accessed and downloaded from <https://datasets.imdbws.com/>. The data is refreshed daily.' Under the heading 'IMDb Dataset Details', it explains: 'Each dataset is contained in a gzipped, tab-separated-values (TSV) formatted file in the UTF-8 character set. The first line in each file contains headers that describe what is in each column. A \"\n\" is used to denote that a particular field is missing or null for that title/name. The available datasets are as follows:'. To the right of the main content area is an advertisement for xfinity mobile, which says 'Switch and save up to \$400 a year on your wireless bill.' and includes a 'Learn more' button and an 'ad feedback' link.

# IMDb Datasets

- **IMDb Dataset Details**

Each dataset is contained in a gzipped, tab-separated-values (TSV) formatted file in the UTF-8 character set. The first line in each file contains headers that describe what is in each column. A '\N' is used to denote that a particular field is missing or null for that title/name.

The available datasets are as follows:

- **title.akas.tsv.gz** - Contains information for titles.
- **title.basics.tsv.gz** - Contains information for titles.
- **title.crew.tsv.gz** – Contains the director and writer information for all the titles in IMDb.
- **title.episode.tsv.gz** – Contains the tv episode information.
- **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 information for names.

# IMDb Datasets: title.akas

**title.akas.tsv.gz** - Contains the following information for titles:

- titleId (string) - a tconst, an alphanumeric unique identifier of the title
- ordering (integer) – a number to uniquely identify rows for a given titleId
- title (string) – the localized title
- region (string) - the region for this version of the title
- language (string) - the language of the title
- types (array) - Enumerated set of attributes for this alternative title. One or more of the following: "alternative", "dvd", "festival", "tv", "video", "working", "original", "imdbDisplay". New values may be added in the future without warning
- attributes (array) - Additional terms to describe this alternative title, not enumerated

# IMDb Datasets: title.basics

**title.basics.tsv.gz** - Contains the following information for titles:

- tconst (string) - alphanumeric unique identifier of the title
- titleType (string) – the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc)
- primaryTitle (string) – the more popular title / the title used by the filmmakers on promotional materials at the point of release
- originalTitle (string) - original title, in the original language
- isAdult (boolean) - 0: non-adult title; 1: adult title
- startYear (YYYY) – represents the release year of a title. In the case of TV Series, it is the series start year
- endYear (YYYY) – TV Series end year. ‘\N’ for all other title types
- runtimeMinutes – primary runtime of the title, in minutes
- genres (string array) – includes up to three genres associated with the title



# IMDb Datasets: title.crew

- **title.crew.tsv.gz** – Contains the director and writer information for all the titles in IMDb. Fields include:
  - tconst (string) - alphanumeric unique identifier of the title
  - directors (array of nconsts) - director(s) of the given title
  - writers (array of nconsts) – writer(s) of the given title

# IMDb Datasets: title.episode

- **title.episode.tsv.gz** – Contains the tv episode information. Fields include:
  - tconst (string) - alphanumeric identifier of episode
  - parentTconst (string) - alphanumeric identifier of the parent TV Series
  - seasonNumber (integer) – season number the episode belongs to
  - episodeNumber (integer) – episode number of the tconst in the TV series

# IMDb Datasets: title.principals

- **title.principals.tsv.gz** – Contains the principal cast/crew for titles
  - tconst (string) - alphanumeric unique identifier of the title
  - ordering (integer) – a number to uniquely identify rows for a given titleId
  - nconst (string) - alphanumeric unique identifier of the name/person
  - category (string) - the category of job that person was in
  - job (string) - the specific job title if applicable, else '\N'
  - characters (string) - the name of the character played if applicable, else '\N'

# IMDb Datasets: title.ratings

- **title.ratings.tsv.gz** – Contains the IMDb rating and votes information for titles
  - tconst (string) - alphanumeric unique identifier of the title
  - averageRating – weighted average of all the individual user ratings
  - numVotes - number of votes the title has received

# IMDb Datasets: name.basics

- **name.basics.tsv.gz** – Contains the following information for names:
  - `nconst` (string) - alphanumeric unique identifier of the name/person
  - `primaryName` (string)– name by which the person is most often credited
  - `birthYear` – in YYYY format
  - `deathYear` – in YYYY format if applicable, else '\N'
  - `primaryProfession` (array of strings)– the top-3 professions of the person
  - `knownForTitles` (array of tconsts) – titles the person is known for

# IMDb Datasets

Table Name	Table Row Count	Source
stg_imdb_name_basics	10,248,845	name_basics.tsv
stg_imdb_name_basics_knownForTitles	16,149,432	name_basics.tsv
stg_imdb_name_basics_primaryProfession	11,215,843	name_basics.tsv
stg_imdb_title_akas	21,614,617	title_akas.tsv
stg_imdb_title_basics	7,017,458	title_basics.tsv
stg_imdb_title_basics_genres	10,593,700	title_basics.tsv
stg_imdb_title_crew	6,704,790	title_crew.tsv
stg_imdb_title_crew_directors	5,170,628	title_crew.tsv
stg_imdb_title_crew_writers	7,916,702	title_crew.tsv
stg_imdb_title_episode	4,764,382	title_episode.tsv
stg_imdb_title_principals	38,699,152	title_principals.tsv
stg_imdb_title_ratings	1,023,897	title_ratings.tsv

Staging tables for  
IMDb core dataset



# IMDb Datasets: ISO Datasets

- Countries
  - countries\_iso - all.tsv
- Languages
  - language-codes-iso.tsv

Table Name	Table Row Count	Source
stg_iso_country	249	countries_iso - all.tsv
stg_iso_language	486	language-codes-iso.tsv

# Movie Lens Data

This dataset (ml-25m) describes 5-star rating and free-text tagging activity from [MovieLens](http://movielens.org), a movie recommendation service. It contains 25,000,095 ratings and 1,093,360 tag applications across 62,423 movies. These data were created by 162,541 users between January 09, 1995 and November 21, 2019. This [dataset](#) was generated on November 21, 2019.

Description: MovieLens\_README.txt

- Movies Data File Structure (MovieLens\_movies.csv)
- Ratings Data File Structure (MovieLens\_ratings.csv)
- Tags Data File Structure (MovieLens\_tags.csv)
- Links Data File Structure (MovieLens\_links.csv)
- Tag Genome (MovieLens\_genome-scores.csv and MovieLens\_genome-tags.csv)

# Movie Lens Data

## Notes:

- Links table cross-map movie lens id with IMDb ids for titles
- **Timestamp is Unix Epoch time**

Table Name	Table Row Count	Source
stg_ml_genome_scores	15,584,448	genome-scores.csv
stg_ml_genome_tags	1,128	genome-tags.csv
stg_ml_links	62,423	links.csv
stg_ml_movies	62,423	movies.csv
stg_ml_ratings	25,000,095	ratings.csv
stg_ml_tags	1,093,360	tags.csv

# IMDb Datasets: Box Office Revenues

- World-Wide Box Office All Time Top 1000 Movies
  - imdb\_project - IMDb Mojo Box Office as 2020-08-02.tsv
  - Note: This is cut & paste from site not a downloaded data set
- There are several titles in this list that do not match the IMDb core dataset
  - You need to identify in reject table
  - determine title that matches
  - add that title to corrected column
  - update target table

TableName	Table Row Count	Comment
stg_box_office_worldwide	994	Before correction
stg_box_office_worldwide	1,000	After data cleansing
stg_box_office_worldwide_reject	6	

# IMDb Project Deliverables

- Ingest initial set of tsv or csv files into staging tables
- Perform data consistency & cleansing processes
- Add supplemental data to model
- Design and create BI visualizations answering business questions





- Data Models





# SQL Script – Microsoft SQL Server

- Create IMDb\_dev & IMDb \_TST databases
- The SQL scripts for each “schema”:
  - Project imdb - STG schema tables.sql
  - Project imdb - INT schema tables.sql
  - Project imdb - BI schema tables.sql
- Create all these tables in each of the databases above

# IMDb Dimensional Model

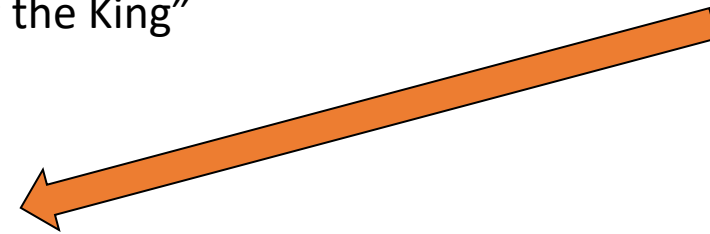
The dimensional model contains:

- Basic information about people in cast, crew, writers and directors
- Basic titles information
- Enhanced title information such as aliases, languages, countries
- Director and writer information for all the titles
- TV episodes information
- Principal cast/crew for titles
- Title Ratings
- Box Office revenue
- Movie franchises
- Movie brands

# IMDb Dimensional Model

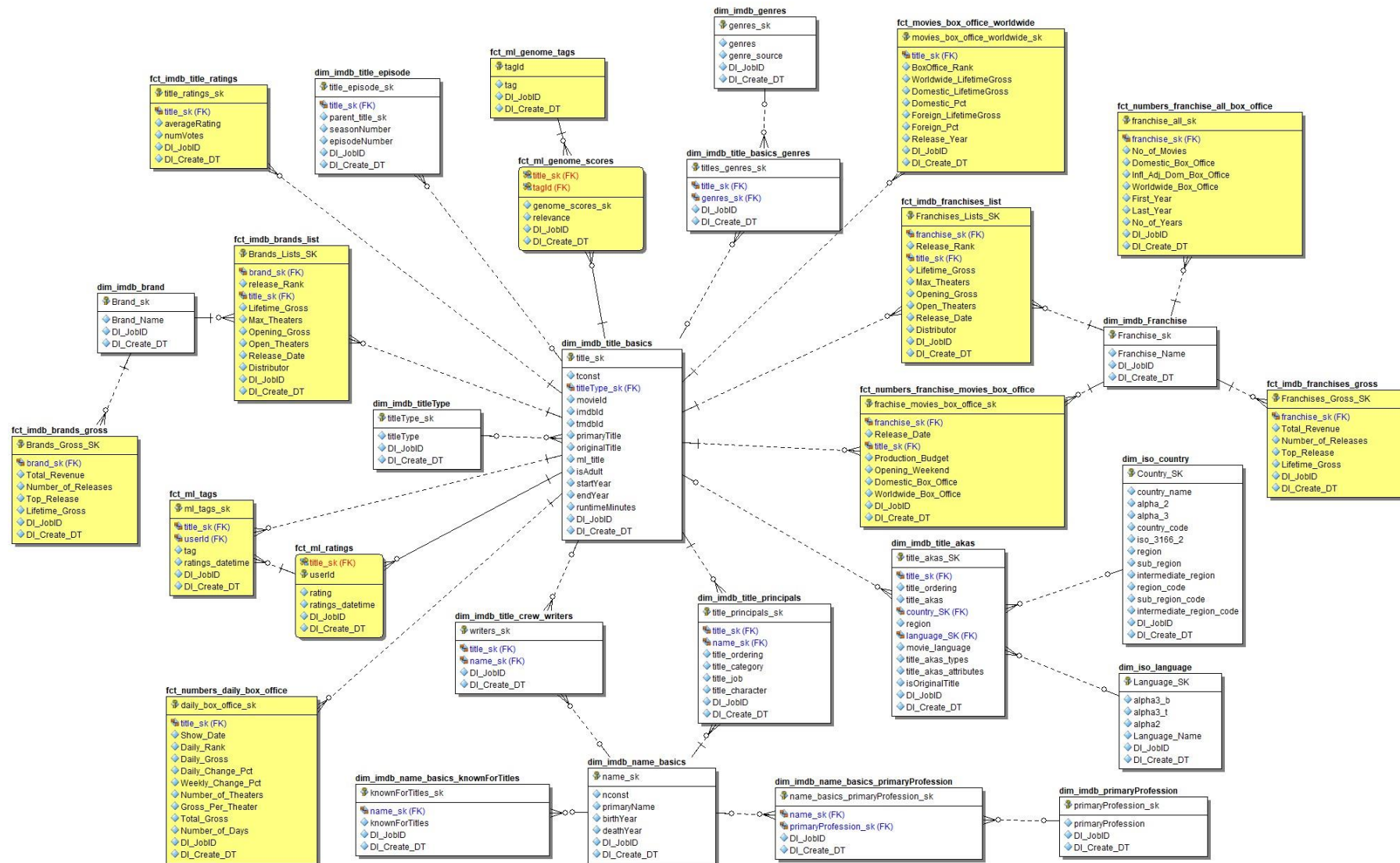
- These is also data about:
  - Countries
  - Languages
  - Movie Genres
  - Job Categories
  - Types of Titles
  - Franchises
  - Brands
- Note: There should be NO repeating groups in STG tables

- All titles and names should have web urls as attributes in dimensional model
  - Title: <https://www.imdb.com/title/> + tconst
    - Example:
      - tconst: "tt0167260"
      - primaryTitle: "The Lord of the Rings: The Return of the King"
      - <https://www.imdb.com/title/tt0167260/>
  - Person: <https://www.imdb.com/name/> + nconst
    - Example
      - nconst: "nm0000949"
      - primaryName = 'Cate Blanchett'
      - <https://www.imdb.com/name/nm0000949/>



# IMDb Dimensional Model

## IMDb Project Dimensional Model



# IMDb Dimensional Model

TableName	Table Row Count
dim_imdb_name_basics	10,248,845
dim_imdb_name_basics_knownForTitles	16,141,204
dim_imdb_name_basics_knownForTitles_rejects	8,228
dim_imdb_name_basics_primaryProfession	11,215,843
dim_imdb_primaryProfession	40
dim_imdb_title_akas	21,597,680
dim_imdb_title_akas_rejects	16,937
dim_imdb_title_basics	7,017,458
dim_imdb_title_basics_genres	10,593,700
dim_imdb_title_crew_directors	5,162,713
dim_imdb_title_crew_directors_rejects	7,915
dim_imdb_title_crew_writers	7,900,554
dim_imdb_title_crew_writers_rejects	16,148
dim_imdb_title_episode	4,761,250
dim_imdb_title_episode_rejects	3,132
dim_imdb_title_principals	38,598,651
dim_imdb_title_principals_rejects	100,501
dim_imdb_titleType	10
dim_iso_country	249
dim_iso_language	486
fct_movies_box_office_worldwide	1,000



# IMDb BI Schema

TableName
bi_top1k_imdb_genres
bi_top1k_imdb_job_category
bi_top1k_imdb_name_basics
bi_top1k_imdb_name_basics_knownForTitles
bi_top1k_imdb_name_basics_primaryProfession
bi_top1k_imdb_primaryProfession
bi_top1k_imdb_title_basics
bi_top1k_imdb_title_basics_genres
bi_top1k_imdb_title_crew_directors
bi_top1k_imdb_title_crew_writers
bi_top1k_imdb_title_principals
bi_top1k_imdb_title_ratings
bi_top1k_imdb_titleType
bi_top1k_iso_country
bi_top1k_iso_language
bi_top1k_ml_genome_scores
bi_top1k_ml_genome_tags
bi_top1k_ml_ratings
bi_top1k_ml_tags
bi_top1k_movies_box_office_worldwide
bi_top1k_numbers_daily_box_office
bi_top1k_numbers_franchise_all_box_office
bi_top1k_numbers_franchise_movies_box_office

- BI Schema # 1
  - Load data based on the Box Office top 1000 movies by revenue
  - Note: This list will require some “manual” data cleansing in order to get the right 1000 movies map to titles and names

# IMDb: Data Integration

- Data Integration
  - Load staging tables
  - Load dimensional model tables
  - Load BI schema tables



# Data Integration – Staging Tables

- Load staging tables
  - Null values in ingestion files need to result in SQL Server column Nulls
  - Data type conversions
- Data Integration Standards
  - All jobs must use Job Statistics Processing Joblets
  - All connections between Talend components need to be labeled, i.e. no row1, row2, etc.
  - Only use the columns needed when ingesting data

- Business Intelligence





# IMDb

- Create dashboards to be able to track entities in dimensional model such as movies, TV episodes and other titles with the people involved with associated revenue and ratings
- Use Microsoft Power BI for all BI
- Use Tableau for selected analysis
  - Analysis of Box Office data

# IMDb - Movies

- Rank the top 100 movies by all time worldwide gross revenue
- For above have report listing various attributes such as release date, running time, rankings, revenue related data, etc.
- Create a dashboard where a top 100 movies is selected the following information is provided:
  - Actors & actresses
  - Writers
  - Directors
  - Genre



# IMDb: Titles

- Rank the top 25 titles by type of title (side panel)
- The ways to rank the title above:
  - IMDb number of votes
  - IMDb rating but use a filter that uses a threshold (minimum number of) votes, i.e. 1M or 100k. The threshold will vary based on type of title
- Type of title:
  - movie
  - short
  - tvEpisode
  - tvMiniSeries
  - tvMovie
  - tvSeries
  - tvShort
  - tvSpecial
  - video
  - videoGame

# Movie Ratings by Movie Lens

- Rank Movies by Movie Lens Ratings the top 25 movies titles in each Movie Lens genre
  - Also list the IMDb worldwide gross (if available) and IMDb ratings

# IMDb - Movies

- Provide a listing various of title attributes such as release date, running time, rankings, revenue related data, category, genre, type of title etc. for selected
  - Actors & actresses
  - Writers
  - Directors
- People to select:
  - John Cusack
  - Ana de Armas
  - Rian Johnson
  - Daisy Ridley
  - Samuel L. Jackson
  - J.J. Abrams
  - Kathryn Bigelow
  - Nicolas Cage
  - Scarlett Johansson
  - Dwayne Johnson
  - Emilia Clarke
  - Woody Harrelson
  - Idris Elba
  - Sean Connery
  - Gal Gadot

# Deliverables

What to upload:

- Talend
  - export of all your jobs with dependencies
  - Screenshots of your jobs' workflows
- BI
  - PowerBI pbix files
  - Tableau twb files
  - Screenshots of dashboards



# Deliverables

## Two online sessions:

- Session with TAs where you will run the complete load from source files to dimensional schema
  - Completeness of data integration
  - Total time to run
  - Table row counts per table
- Team Presentation
  - Review of data integration
    - Workflow, Transformations & Rejects
  - Review of BI
    - Answering ?s in Power BI
    - Displaying visualizations in Tableau
    - Any business analysis you feel tells a story

