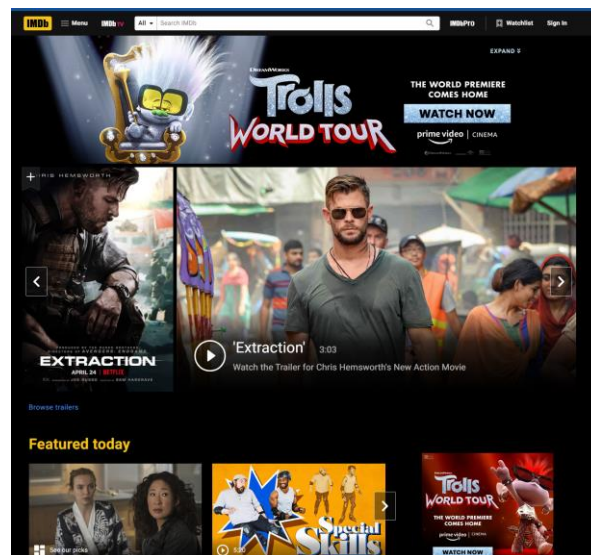# IMDb Data & Analysis

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

BUSINESS INTELLIGENCE GUIDEBOOK

From Data Integration to Analytics

RICK SHERMAN

FOREWORD BY **CLAUDIA IMHOFF**
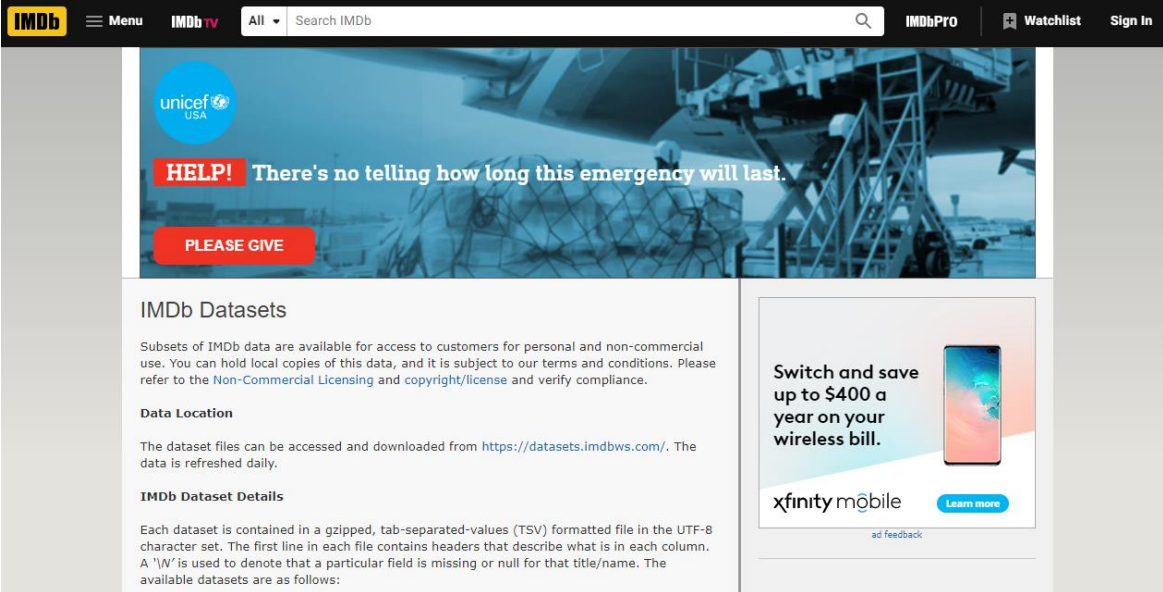PRESIDENT OF INTELLIGENT SOLUTIONS, INC.

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.

ATHENA
IT SOLUTIONS

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.

ATHENA
IT SOLUTIONS

IMDb Datasets

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

ATHENA
IT SOLUTIONS

IMDb Datasets

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

ATHENA
IT SOLUTIONS

IMDb Datasets

- **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
- **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

ATHENA
IT SOLUTIONS

IMDb Datasets

- **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.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

- **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

ATHENA
IT SOLUTIONS

IMDb Datasets – ISO Datasets

- Countries
  - countries_iso - all.tsv

- Languages
  - language-codes-iso.tsv

IMDb Datasets – Box Office Revenues

- World Wide Box Office All Time Top 1000 Movies
  - World Wide Box Office All Time Top 1000.tsv

- Top Movie Franchises
  - IMDb BoxOfficeMojo - Franchises (US & Canada).tsv
  - IMDb BoxOfficeMojo - Franchise_ Marvel Cinematic Universe.tsv

- Top Movie Brands
  - IMDb BoxOfficeMojo - Brands (US & Canada).tsv
  - IMDb BoxOfficeMojo - Brand_ Marvel Comics.tsv

ATHENA
IT SOLUTIONS

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)

ATHENA
IT SOLUTIONS

Movie Lens Data

SQL script for Stage tables:

- stg_ml_tables.sql

Notes:

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

TableName	Table_Rows
stg_ml_genome_scores	15,584,448
stg_ml_genome_tags	1,128
stg_ml_links	62,423
stg_ml_movies	62,423
stg_ml_ratings	25,000,095
stg_ml_tags	1,093,360

ATHENA
IT SOLUTIONS

IMDb Project Deliverables

- Ingest initial set of tsv or csv files into staging tables

- Design and load dimensional model for above data

- Perform data consistency & cleansing processes

- Add supplemental data to model

- Design and create BI visualizations answering business questions

ATHENA
IT SOLUTIONS

IMDb

- Data Modeling

- Two common source scenarios:
  - Entity-Relationship (ER) Model in complex enterprise application
  - Denormalized (or flattened) data
    - Spreadsheets
    - BI Application design

ATHENA
IT SOLUTIONS

SQL Script – Microsoft SQL Server

- Create IMDb & IMDb _TST databases

- Create tables for Staging using SQL script on OneDrive class drive:
  - Ingestion or Staging Tables for Microsoft SQL Server

- SQL Server scripts
  - stg imdb tables - core tables.sql
  - stg imdb tables expanded part 1.sql
  - stg imdb tables expanded part 2.sql

ATHENA
IT SOLUTIONS

IMDb Staging Tables

TableName	Table_Rows
stg_imdb_name_basics	10,008,408
stg_imdb_name_basics_knownForTitles	21,575,298
stg_imdb_name_basics_primaryProfession	10,986,452
stg_imdb_title_akas	21,614,617
stg_imdb_title_basics	6,704,790
stg_imdb_title_basics_genres	10,158,625
stg_imdb_title_crew	6,704,790
stg_imdb_title_crew_directors	5,170,628
stg_imdb_title_crew_writers	7,916,702
stg_imdb_title_episode	4,764,382
stg_imdb_title_principals	38,699,152
stg_imdb_title_ratings	1,023,897
stg_iso_country	249
stg_iso_language	486
stg_box_office_worldwide	1000
stg_imdb_brands_gross	44
stg_imdb_brands_list	59
stg_imdb_franchises_gross	289
stg_imdb_franchises_list	35

ATHENA
IT SOLUTIONS

IMDb Dimensional Model

Need to be able to model:

- 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

- Also need to create dimensional tables for:
    - Country
    - Languages
    - Movie Genres
    - Job Categories
    - Types of Titles

ATHENA
IT SOLUTIONS

IMDb Dimensional Model

- 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/title/ + nconst
    - Example
      - nconst: "nm0000949"
      - primaryName = 'Cate Blanchett'
      - https://www.imdb.com/name/nm0000949/

ATHENA
IT SOLUTIONS

IMDb

- ## Data Integration
  - Load staging tables
  - Load dimensional model tables

ATHENA
IT SOLUTIONS

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

ATHENA
IT SOLUTIONS

Data Integration

IMDb

- 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
    - TBD

ATHENA
IT SOLUTIONS

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

ATHENA
IT SOLUTIONS

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

ATHENA
IT SOLUTIONS

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

ATHENA
IT SOLUTIONS

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

IMDb

- Business questions