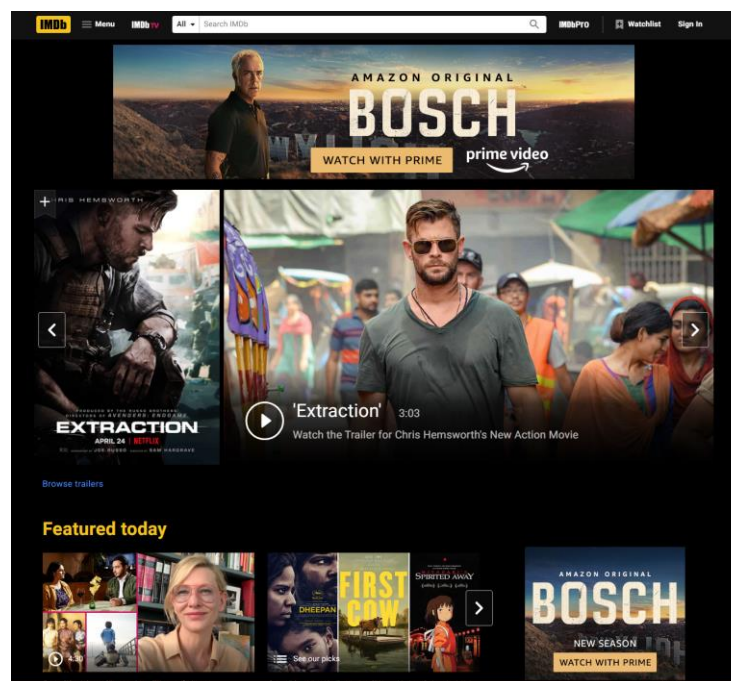


IMDb Data & Analysis

IMDb

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



BUSINESS INTELLIGENCE GUIDEBOOK

From Data Integration to Analytics

MK
MORGAN KAUFMANN

RICK SHERMAN

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

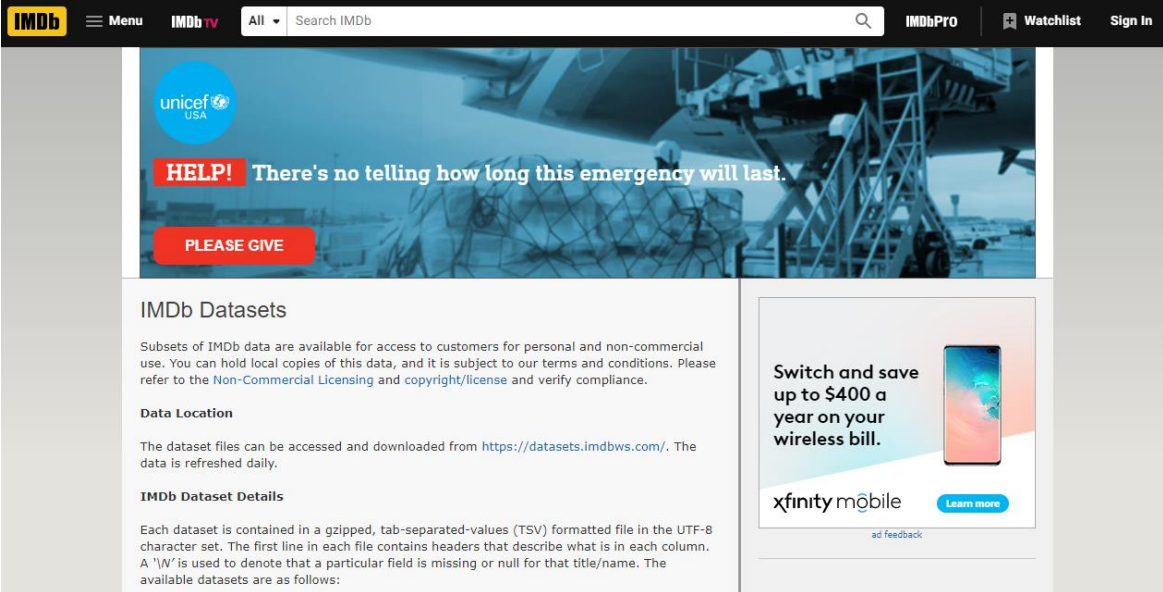
IMDb Datasets

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.



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 says: '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 with the text 'Switch and save up to \$400 a year on your wireless bill.' and a 'Learn more' button.

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.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.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.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

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

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 – aggregate data for all franchises
 - IMDb BoxOfficeMojo - Franchise_ Marvel Cinematic Universe.tsv – data for one franchise
 - You need to extract & load data for top 20 franchises
- Top Movie Brands
 - IMDb BoxOfficeMojo - Brands (US & Canada).tsv - aggregate data for all brands
 - IMDb BoxOfficeMojo - Brand_ Marvel Comics.tsv – data for one brand
 - You need to extract & load data for top 20 brands

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

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

The Numbers

(Note: Not too excited about using this site but..)

- Obtaining daily box office data on franchises & their movies by cut & past
 - [Franchises Domestic Box Office](#)
 - [Box Office History for Marvel Cinematic Universe Movies](#)
 - [The Avengers \(2012\)](#)
- Files
 - The Numbers - Domestic Box Office - Franchises.tsv
 - The Numbers - Domestic Box Office - Franchises - Marvel Cinematic Universe.tsv
 - The Numbers - Domestic Box Office Daily - The Avengers.tsv
- SQL script: The Number - stage tables.sql

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



- Data Modeling



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

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

IMDb Dimensional Model

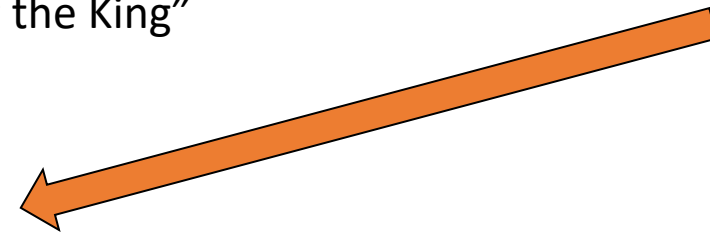
Need to be able to model:

- 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

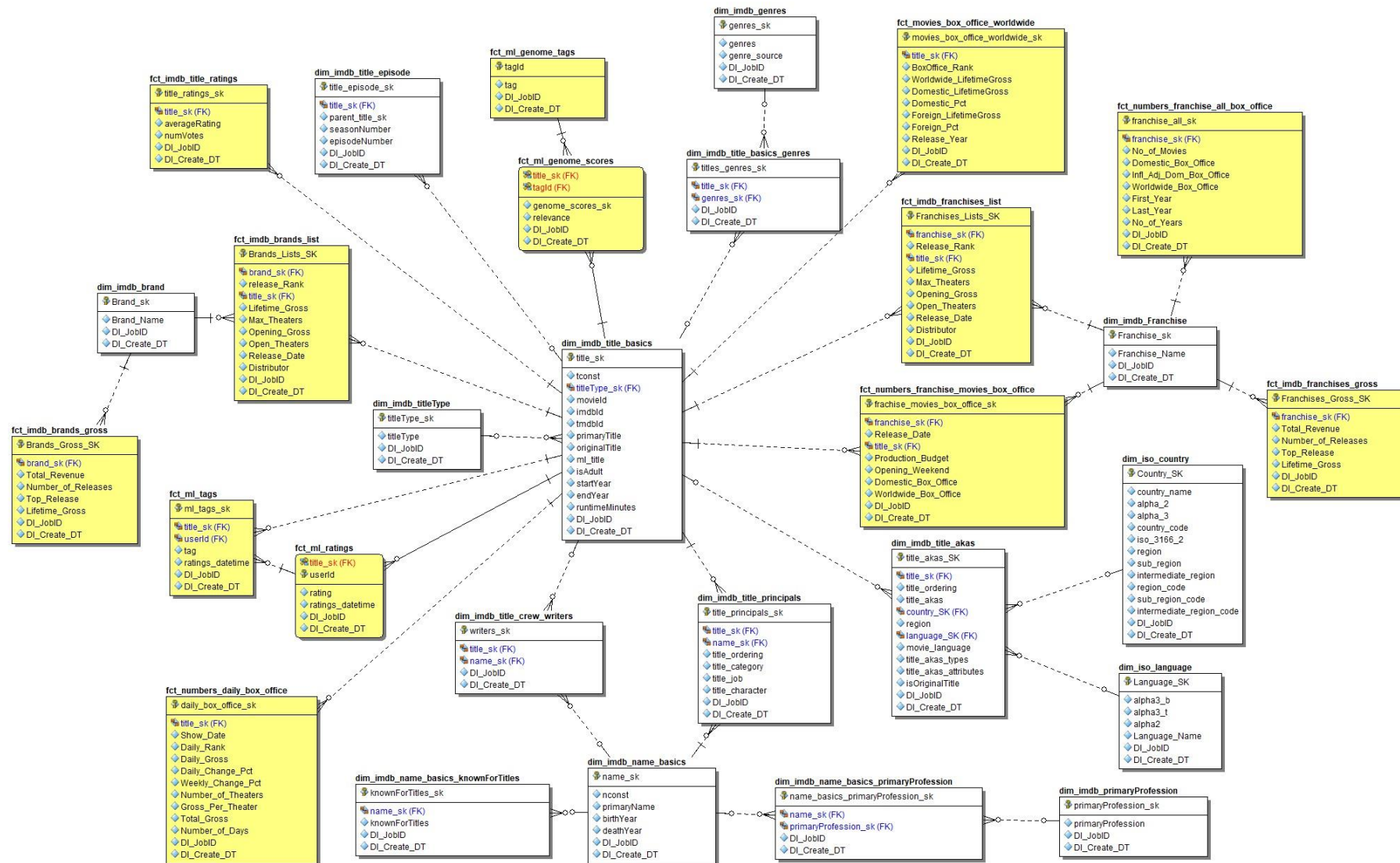
- Also need to create dimensional tables for:
 - Country
 - Languages
 - Movie Genres
 - Job Categories
 - Types of Titles
 - Franchises
 - Brands
- No repeating groups

- 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

SQL script: DDL for IMDb Project Dimensional Model.sql



IMDb Dimensional Model

Dimensions

- dim_imdb_brand
- dim_imdb_Franchise
- dim_imdb_genres
- dim_imdb_genres_ml_rejects
- dim_imdb_job_category
- dim_imdb_name_basics
- dim_imdb_name_basics_knownForTitles
- dim_imdb_name_basics_knownForTitles_rejects
- dim_imdb_name_basics_primaryProfession
- dim_imdb_primaryProfession
- dim_imdb_title_akas
- dim_imdb_title_akas_rejects
- dim_imdb_title_basics
- dim_imdb_title_basics_genres
- dim_imdb_title_basics_ml_rejects
- dim_imdb_title_crew_directors
- dim_imdb_title_crew_directors_rejects
- dim_imdb_title_crew_writers
- dim_imdb_title_crew_writers_rejects
- dim_imdb_title_episode
- dim_imdb_title_episode_rejects
- dim_imdb_title_principals
- dim_imdb_title_principals_rejects
- dim_imdb_titleType
- dim_imdbtitle_basics_rejects_ml
- dim_iso_country
- dim_iso_language

Facts

- fct_imdb_brands_gross
- fct_imdb_brands_list
- fct_imdb_franchises_gross
- fct_imdb_franchises_list
- fct_imdb_title_ratings
- fct_ml_genome_scores
- fct_ml_genome_scores_rejects
- fct_ml_genome_tags
- fct_ml_ratings
- fct_ml_ratings_rejects
- fct_ml_tags
- fct_ml_tags_rejects
- fct_movies_box_office_worldwide
- fct_numbers_daily_box_office
- fct_numbers_franchise_all_box_office
- fct_numbers_franchise_movies_box_office

IMDb Dimensional Model

TableName	TableRows
dim_imdb_brand	44
dim_imdb_Franchise	289
dim_imdb_genres	48
dim_imdb_genres_ml_rejects	138
dim_imdb_job_category	12
dim_imdb_name_basics	10,008,408
dim_imdb_name_basics_knownForTitles	21,565,259
dim_imdb_name_basics_knownForTitles_rejects	10,039
dim_imdb_name_basics_primaryProfession	10,986,452
dim_imdb_primaryProfession	40
dim_imdb_title_akas	21,609,393
dim_imdb_title_akas_rejects	5,224
dim_imdb_title_basics	6,704,790
dim_imdb_title_basics_genres	10,270,794
dim_imdb_title_basics_ml_rejects	78
dim_imdb_title_crew_directors	5,170,597
dim_imdb_title_crew_directors_rejects	31
dim_imdb_title_crew_writers	7,916,672
dim_imdb_title_crew_writers_rejects	30
dim_imdb_title_episode	4,764,372
dim_imdb_title_episode_rejects	10
dim_imdb_title_principals	38,676,545
dim_imdb_title_principals_rejects	22,607
dim_imdb_titleType	10
dim_imdbtitle_basics_rejects_ml	164,058
dim_iso_country	249
dim_iso_language	486

TableName	TableRows
fct_imdb_brands_gross	44
fct_imdb_franchises_gross	289
fct_imdb_franchises_list	281
fct_imdb_title_ratings	1,023,897
fct_ml_genome_scores	15,560,760
fct_ml_genome_scores_rejects	23,688
fct_ml_genome_tags	1,128
fct_ml_ratings	24,983,941
fct_ml_ratings_rejects	16,154
fct_ml_tags	1,092,751
fct_ml_tags_rejects	609

- Data Integration

- Load staging tables
- Load dimensional model 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

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

