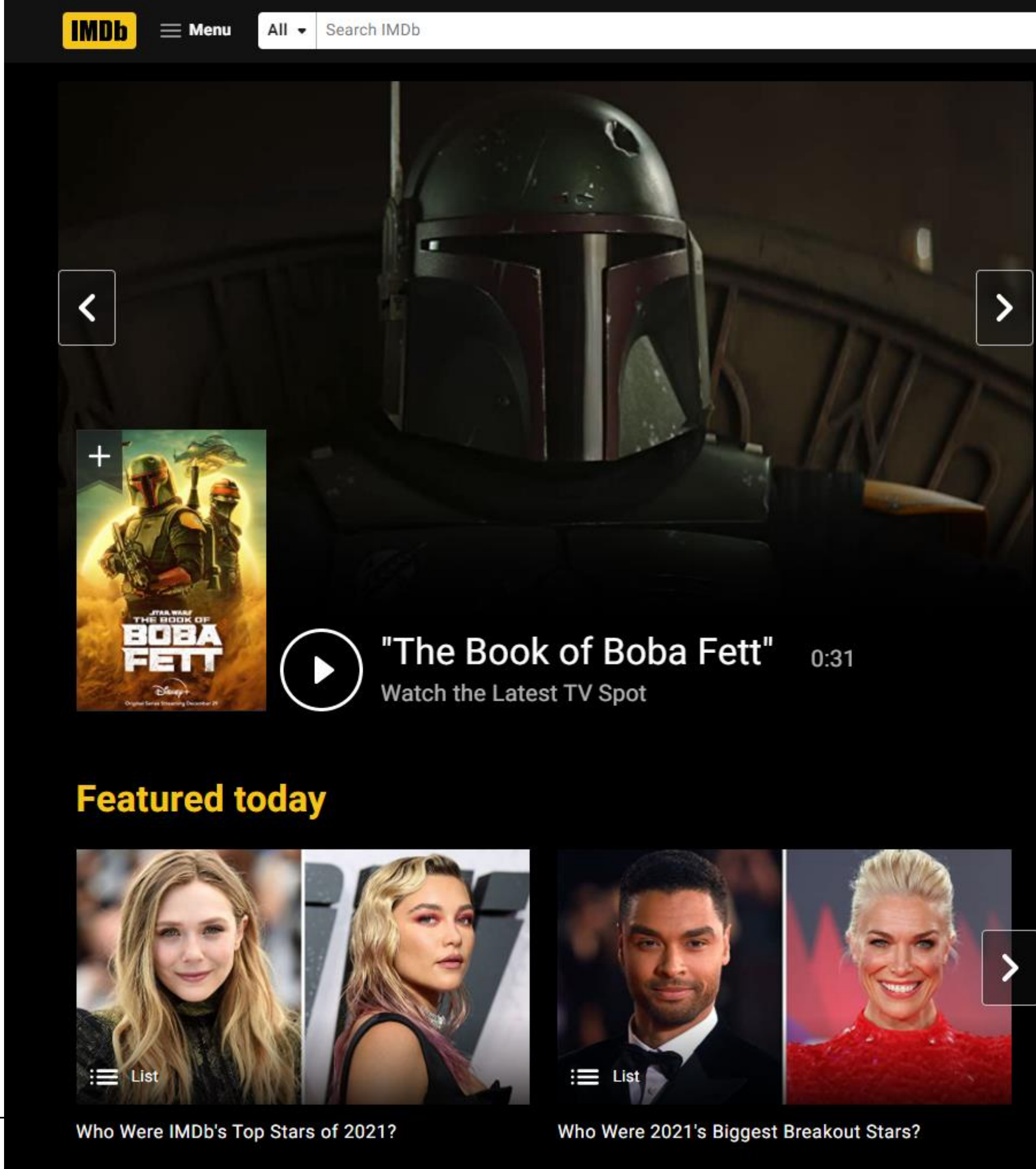
UPDATED

2022-04-25
7:30am

IMDb Data & Analysis Project: Staging

Rick Sherman



"The Book of Boba Fett" 0:31
Watch the Latest TV Spot

Featured today

Who Were IMDb's Top Stars of 2021?

Who Were 2021's Biggest Breakout Stars?

IMDb Project - Staging

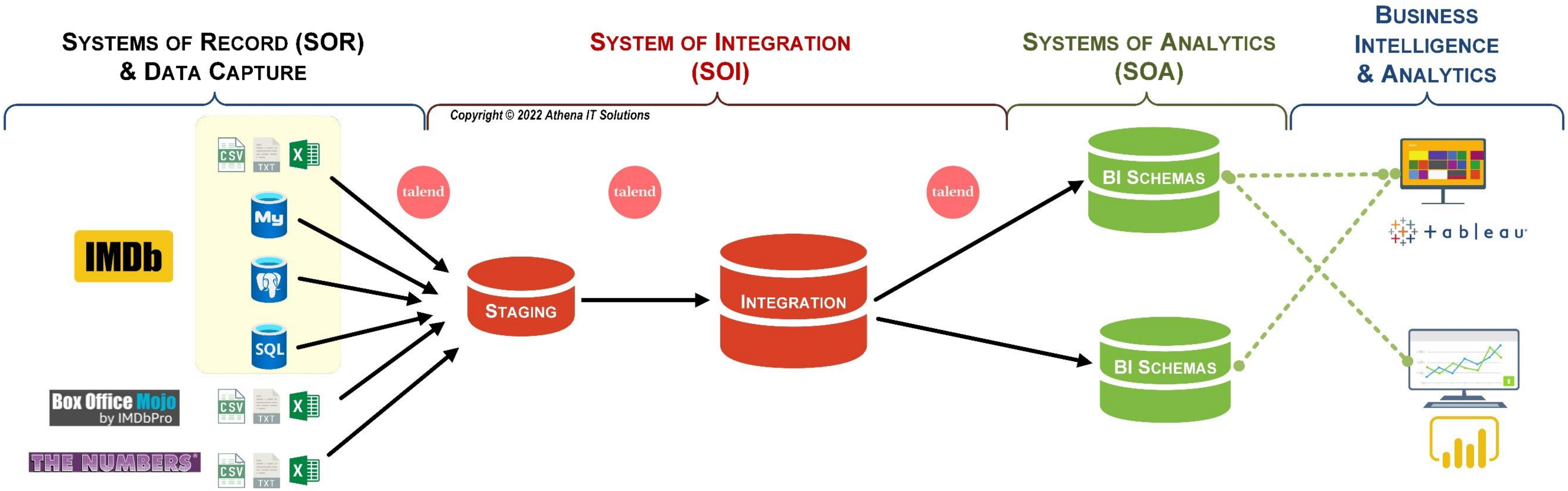
ADA subproject (Core):

- Load Data into ADA schemas in SQL Server using Talend
  - IMDb basic data split by title category
  - IMDb lists, The Numbers and IMDb Pro

Staging only subproject:

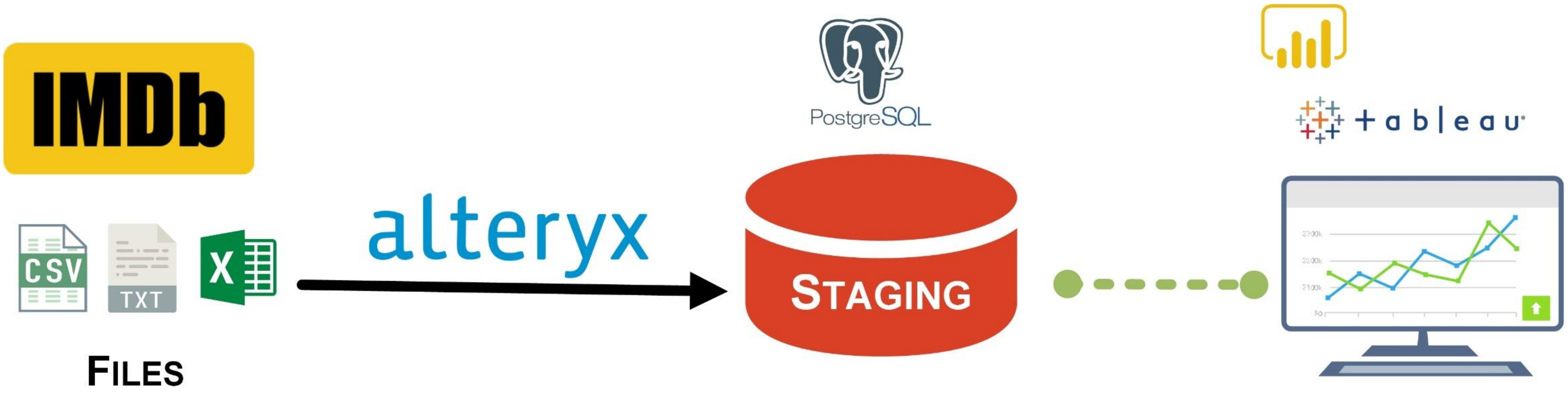
- Load IMDb dataset into Stage tables in PostgreSQL using Alteryx

IMDb Project Architecture – ADA subproject



- Implementing an Analytica Data Architecture (ADA)
  - Staging – load data from data sources
  - Integration – load into a dimensional model that provides data consistency
  - BI Schemas – create one or more BI schemas tailored to specific types of analysis
    - San Francisco Film Locations
    - Examples: Top 1000 movies by revenue or Top 100 TV shows by rankings, both with associated data

IMDb Project Architecture: Staging Only Subproject



- Deliverables:
  - Perform data profiling
  - Load Staging Tables with data preparation tool
  - Create dashboards enabling selection of a title or person
- Notes:
  - This is NOT used for Integration or BI schema loading

Staging Only Subproject

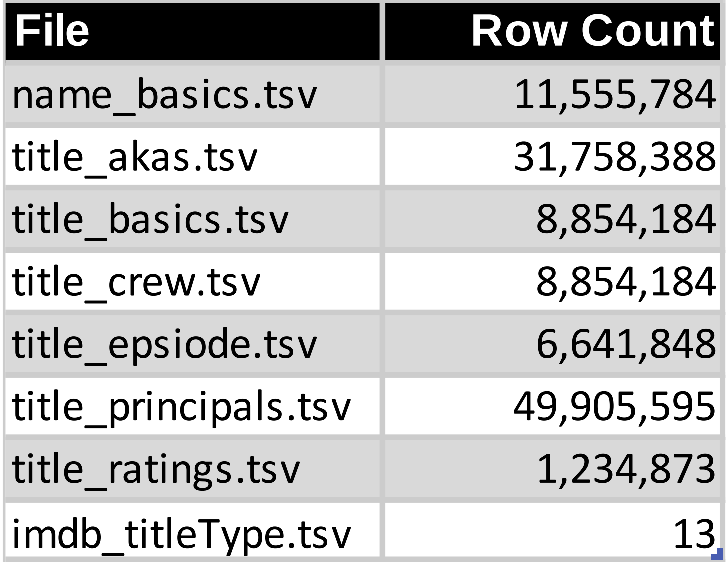
Load IMDb datasets only

Datasets: Staging Only Subproject

UPDATED

\Project - imdb\data_imdb_datasets\imdb_tsv_files

- name_basics.tsv
- title_akas.tsv
- title_basics.tsv
- title_crew.tsv
- title_epsiode.tsv
- title_principals.tsv
- title_ratings.tsv

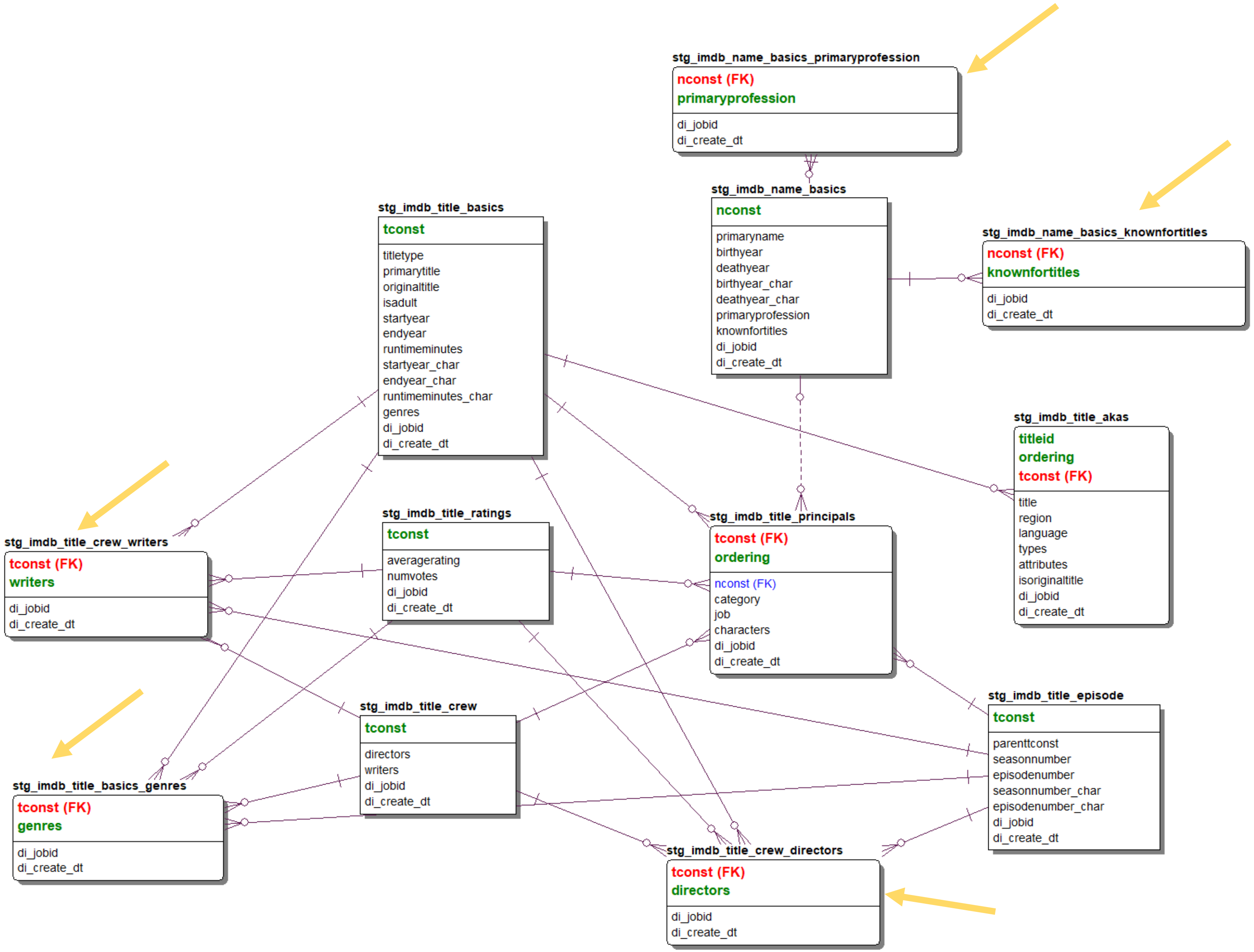


File	Row Count
name_basics.tsv	11,555,784
title_akas.tsv	31,758,388
title_basics.tsv	8,854,184
title_crew.tsv	8,854,184
title_epsiode.tsv	6,641,848
title_principals.tsv	49,905,595
title_ratings.tsv	1,234,873
imdb_titleType.tsv	13

- Staging schema script (PostgreSQL)
  - \Project - imdb\schema_sql
    - imdb stg for Alteryx - PostgreSQL - IMDb datasets only.sql

Stage Data Model: Staging Only Subproject

UPDATED



stg_imdb_name_basics_primaryprofession
nconst (FK)
primaryprofession
di_jobid
di_create_dt

stg_imdb_name_basics
nconst
primaryname
birthyear
deathyear
birthyear_char
deathyear_char
primaryprofession
knownfortitles
di_jobid
di_create_dt

stg_imdb_name_basics_knownfortitles
nconst (FK)
knownfortitles
di_jobid
di_create_dt

stg_imdb_title_basics
tconst
titletype
primarytitle
originaltitle
isadult
startyear
endyear
runtimeminutes
startyear_char
endyear_char
runtimeminutes_char
genres
di_jobid
di_create_dt

stg_imdb_title_akas
titleid
ordering
tconst (FK)
title
region
language
types
attributes
isoriginaltitle
di_jobid
di_create_dt

stg_imdb_title_crew_writers
tconst (FK)
writers
di_jobid
di_create_dt

stg_imdb_title_ratings
tconst
averagerating
numvotes
di_jobid
di_create_dt

stg_imdb_title_principals
tconst (FK)
ordering
nconst (FK)
category
job
characters
di_jobid
di_create_dt

stg_imdb_title_crew
tconst
directors
writers
di_jobid
di_create_dt

stg_imdb_title_episode
tconst
parenttconst
seasonnumber
episodenumber
seasonnumber_char
episodenumber_char
di_jobid
di_create_dt

stg_imdb_title_basics_genres
tconst (FK)
genres
di_jobid
di_create_dt

stg_imdb_title_crew_directors
tconst (FK)
directors
di_jobid
di_create_dt

imdb_stage Tables
stg_imdb_name_basics
stg_imdb_name_basics_knownfortitles
stg_imdb_name_basics_primaryprofession
stg_imdb_title_akas
stg_imdb_title_basics
stg_imdb_title_basics_genres
stg_imdb_title_crew
stg_imdb_title_crew_directors
stg_imdb_title_crew_writers
stg_imdb_title_episode
stg_imdb_title_principals
stg_imdb_title_ratings

Yellow Arrow indicates tables where repeating groups are normalized when ingesting data

Deliverables: Staging Only Subproject

UPDATED

- Loading data into staging tables
- Data Profiling
  - List row counts
- Data Discovery using Power BI dashboards on staging tables
  - Questions to answer - TBD
    - List all information about a title
    - List all information about a person
    - Table & row count

    - TBD: Filters for above

ADA subproject (Core)

Load IMDb datasets and supplemental data

ADA subproject (Core): Data Integration

UPDATED

- Ingest initial datasets into staging schema (ingestion)
  - IMDb core data for Movies
  - IMDb core data for TV
  - IMDb core data for Short (Movies)
  - IMDb core data for Videos & Misc.
  - IMDb Mojo data for Box Office, Brands, Franchises and Genes
  - The Numbers data for movie box office
  - IMDb Lists (NOTE: This will be done in Part 2)
- Integrate data into integration schema (dimensional data model)
  - Dimensions: much of core data is dimensions
  - Facts: examples are box office sales, ratings, budgets, etc.

ADA subproject (Core): Sources – imdb data

UPDATED

- IMDb core data for Movies
  - \Project - imdb\data_imdb_datasets\imdb_src_movies
    - imdb_src_movies.sql (MySQL backup)
- IMDb core data for TV
  - \Project - imdb\data_imdb_datasets\imdb_src_tv
    - imdb_src_tv.bak (SQL Server backup)

ADA subproject (Core):

- IMDb core data for Short (Movies)
  - \Project - imdb\data_imdb_datasets\imdb_src_short
    - dump-imdb_src_short.dump (PostgreSQL backup/dump file)
- IMDb core data for Videos & Misc.
  - \Project - imdb\data_imdb_datasets\imdb_src_videos
    - name_basics_misc.tsv
    - name_basics_video.tsv
    - title_akas_misc.tsv
    - title_akas_video.tsv
    - title_basics_videos.tsv
    - title_crew_video.tsv
    - title_principals_misc.tsv
    - title_principals_video.tsv
    - title_ratings_video.tsv

ADA subproject (Core): Sources – imdb data

NEW

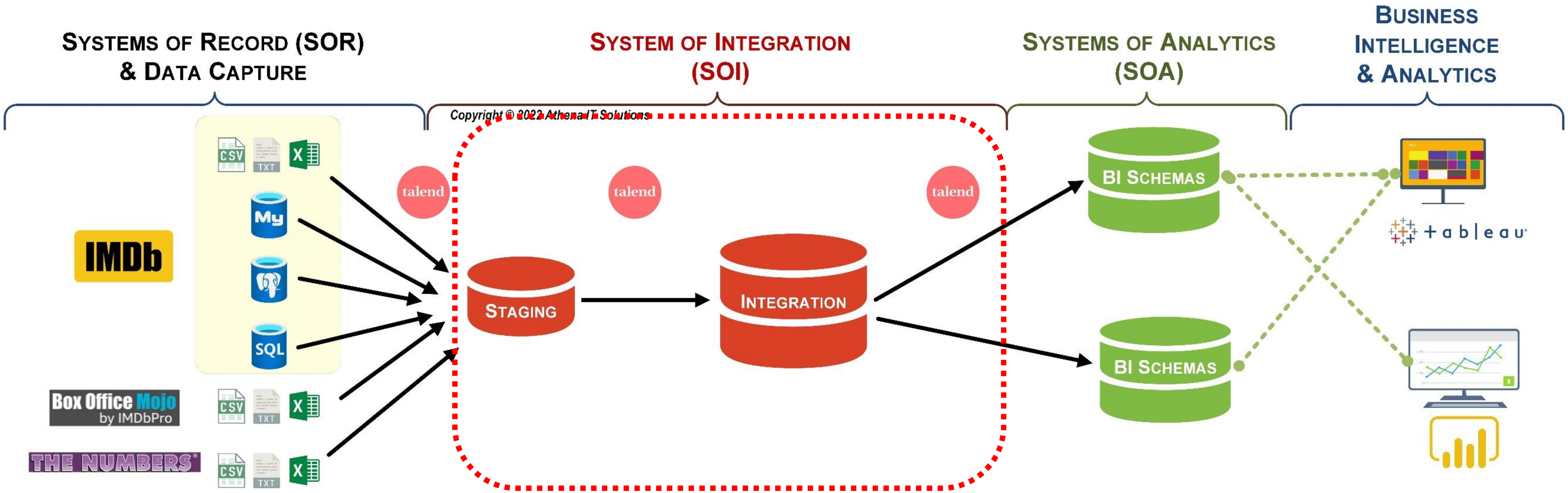
Source	MySQL	SQL Server	PostgreSQL	tsv files	tsv files
file	**movies**	**tv**	**short**	**videos**	**misc**
name_basics	1,454,602	2,117,967	2,004,666	547,616	6,861,340
title_akas	2,632,798	26,703,795	1,771,693	644,879	5,223
title_basics	607,423	7,091,521	865,066	290,174	0
title_crew	607,423	7,091,502	865,066	290,174	0
title_epsiode	0	6,641,848	0	0	0
title_principals	4,283,620	39,646,094	4,375,065	1,599,978	838
title_ratings	277,171	753,531	140,980	63,191	0
imdb_titleType	13	0	13	13	0

ADA subproject (Core): SQL Scripts

- Staging schema script (SQL Server or Azure SQL)
  - \Project - imdb\schema_sql
    - Project imdb - STG schema tables - SQL Server 2022-04-21.sql

- Integration schema script (SQL Server or Azure SQL)
  - \Project - imdb\schema_sql
    - Project imdb - INT schema tables - SQL Server 2022-04-21.sql

ADA subproject (Core): IMDb Project Architecture

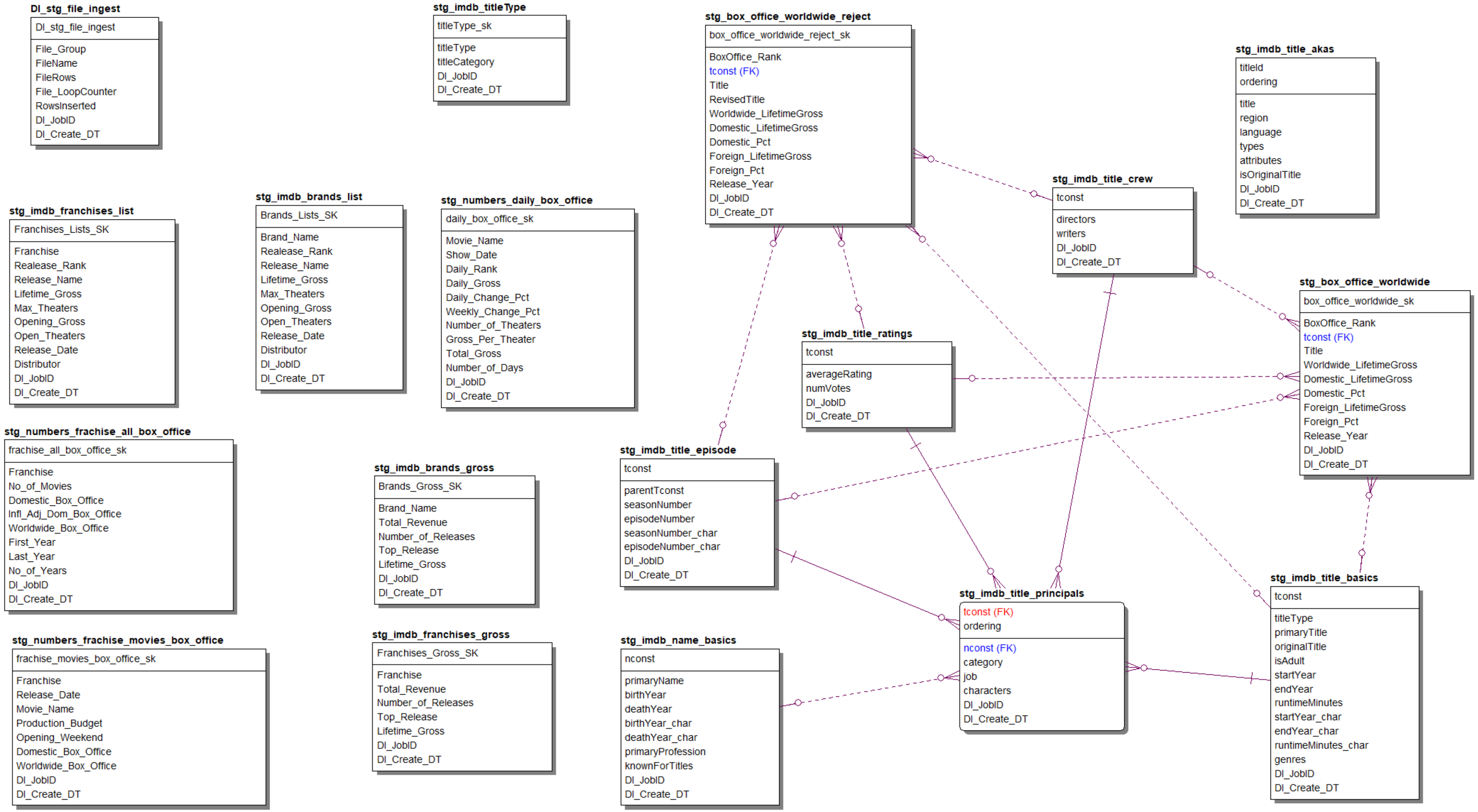


SYSTEMS OF RECORD (SOR) & DATA CAPTURE
SYSTEM OF INTEGRATION (SOI)
SYSTEMS OF ANALYTICS (SOA)
BUSINESS INTELLIGENCE & ANALYTICS
Copyright © 2022 Athena IT Solutions
CSV
TXT
My
SQL
IMDb
Box Office Mojo by IMDbPro
THE NUMBERS
talend
talend
talend
STAGING
INTEGRATION
BI SCHEMAS
BI SCHEMAS
+ a b | e a u

- Implementing an Analytica Data Architecture (ADA)
  - Staging – load data from data sources
  - Integration – load into a dimensional model that provides data consistency
  - BI Schemas – create one or more BI schemas tailored to specific types of analysis
    - San Francisco Film Locations
    - Examples: Top 1000 movies by revenue or Top 100 TV shows by rankings, both with associated data

ADA subproject (Core): Submodel - STG

UPDATED



DI_stg_file_ingest
DI_stg_file_ingest
File_Group
FileName
FileRows
File_LoopCounter
RowsInserted
DI_JobID
DI_Create_DT

stg_imdb_titleType
titleType_sk
titleType
titleCategory
DI_JobID
DI_Create_DT

stg_box_office_worldwide_reject
box_office_worldwide_reject_sk
BoxOffice_Rank
tconst (FK)
Title
RevisedTitle
Worldwide_LifetimeGross
Domestic_LifetimeGross
Domestic_Pct
Foreign_LifetimeGross
Foreign_Pct
Release_Year
DI_JobID
DI_Create_DT

stg_imdb_title_akas
titleId
ordering
title
region
language
types
attributes
isOriginalTitle
DI_JobID
DI_Create_DT

stg_imdb_franchises_list
Franchises_Lists_SK
Franchise
Realease_Rank
Release_Name
Lifetime_Gross
Max_Theaters
Opening_Gross
Open_Theaters
Release_Date
Distributor
DI_JobID
DI_Create_DT

stg_imdb_brands_list
Brands_Lists_SK
Brand_Name
Realease_Rank
Release_Name
Lifetime_Gross
Max_Theaters
Opening_Gross
Open_Theaters
Release_Date
Distributor
DI_JobID
DI_Create_DT

stg_numbers_daily_box_office
daily_box_office_sk
Movie_Name
Show_Date
Daily_Rank
Daily_Gross
Daily_Change_Pct
Weekly_Change_Pct
Number_of_Theaters
Gross_Per_Theater
Total_Gross
Number_of_Days
DI_JobID
DI_Create_DT

stg_imdb_title_crew
tconst
directors
writers
DI_JobID
DI_Create_DT

stg_box_office_worldwide
box_office_worldwide_sk
BoxOffice_Rank
tconst (FK)
Title
Worldwide_LifetimeGross
Domestic_LifetimeGross
Domestic_Pct
Foreign_LifetimeGross
Foreign_Pct
Release_Year
DI_JobID
DI_Create_DT

stg_imdb_title_ratings
tconst
averageRating
numVotes
DI_JobID
DI_Create_DT

stg_numbers_frachise_all_box_office
frachise_all_box_office_sk
Franchise
No_of_Movies
Domestic_Box_Office
Infl_Adj_Dom_Box_Office
Worldwide_Box_Office
First_Year
Last_Year
No_of_Years
DI_JobID
DI_Create_DT

stg_imdb_brands_gross
Brands_Gross_SK
Brand_Name
Total_Revenue
Number_of_Releases
Top_Release
Lifetime_Gross
DI_JobID
DI_Create_DT

stg_imdb_title_episode
tconst
parentTconst
seasonNumber
episodeNumber
seasonNumber_char
episodeNumber_char
DI_JobID
DI_Create_DT

stg_imdb_title_principals
tconst (FK)
ordering
nconst (FK)
category
job
characters
DI_JobID
DI_Create_DT

stg_imdb_title_basics
tconst
titleType
primaryTitle
originalTitle
isAdult
startYear
endYear
runtimeMinutes
startYear_char
endYear_char
runtimeMinutes_char
genres
DI_JobID
DI_Create_DT

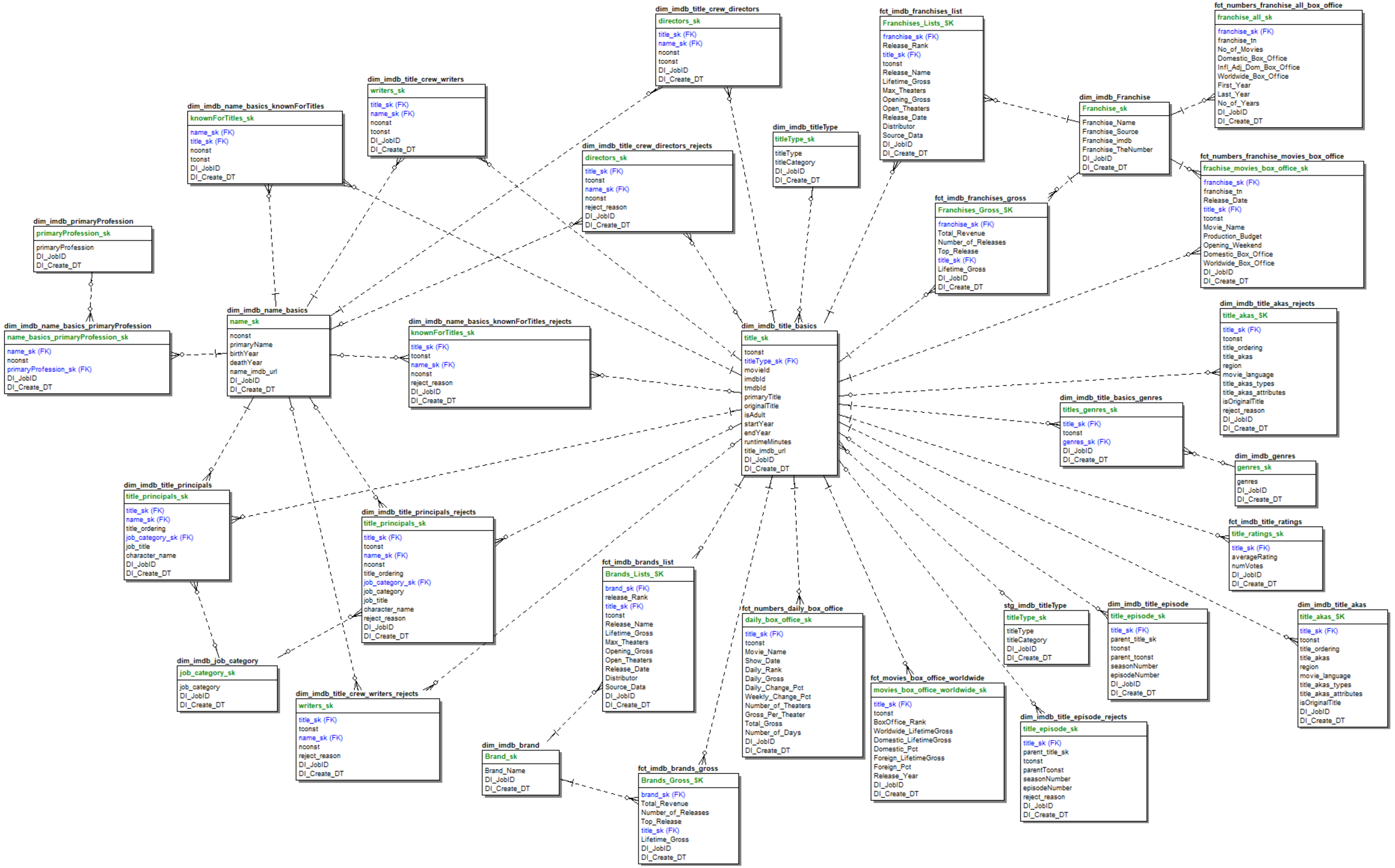
stg_numbers_frachise_movies_box_office
frachise_movies_box_office_sk
Franchise
Release_Date
Movie_Name
Production_Budget
Opening_Weekend
Domestic_Box_Office
Worldwide_Box_Office
DI_JobID
DI_Create_DT

stg_imdb_franchises_gross
Franchises_Gross_SK
Franchise
Total_Revenue
Number_of_Releases
Top_Release
Lifetime_Gross
DI_JobID
DI_Create_DT

stg_imdb_name_basics
nconst
primaryName
birthYear
deathYear
birthYear_char
deathYear_char
primaryProfession
knownForTitles
DI_JobID
DI_Create_DT

ADA subproject (Core): Submodel - INT

UPDATED



dim_imdb_title_crew_directors
directors_sk
title_sk (FK)
name_sk (FK)
nconst
tconst
DI_JobID
DI_Create_DT

fct_imdb_franchises_list
Franchises_Lists_SK
franchise_sk (FK)
Release_Rank
title_sk (FK)
tconst
Release_Name
Lifetime_Gross
Max_Theaters
Opening_Gross
Open_Theaters
Release_Date
Distributor
Source_Data
DI_JobID
DI_Create_DT

fct_numbers_franchise_all_box_office
franchise_all_sk
franchise_sk (FK)
franchise_tn
No_of_Movies
Domestic_Box_Office
Infl_Adj_Dom_Box_Office
Worldwide_Box_Office
First_Year
Last_Year
No_of_Years
DI_JobID
DI_Create_DT

dim_imdb_title_crew_writers
writers_sk
title_sk (FK)
name_sk (FK)
nconst
tconst
DI_JobID
DI_Create_DT

dim_imdb_name_basics_knownForTitles
knownForTitles_sk
name_sk (FK)
title_sk (FK)
nconst
tconst
DI_JobID
DI_Create_DT

dim_imdb_Franchise
Franchise_sk
Franchise_Name
Franchise_Source
Franchise_imdb
Franchise_TheNumber
DI_JobID
DI_Create_DT

dim_imdb_titleType
titleType_sk
titleType
titleCategory
DI_JobID
DI_Create_DT

dim_imdb_title_crew_directors_rejects
directors_sk
title_sk (FK)
tconst
name_sk (FK)
nconst
reject_reason
DI_JobID
DI_Create_DT

fct_numbers_franchise_movies_box_office
frachise_movies_box_office_sk
franchise_sk (FK)
franchise_tn
Release_Date
title_sk (FK)
tconst
Movie_Name
Production_Budget
Opening_Weekend
Domestic_Box_Office
Worldwide_Box_Office
DI_JobID
DI_Create_DT

fct_imdb_franchises_gross
Franchises_Gross_SK
franchise_sk (FK)
Total_Revenue
Number_of_Releases
Top_Release
title_sk (FK)
Lifetime_Gross
DI_JobID
DI_Create_DT

dim_imdb_primaryProfession
primaryProfession_sk
primaryProfession
DI_JobID
DI_Create_DT

dim_imdb_title_akas_rejects
title_akas_SK
title_sk (FK)
tconst
title_ordering
title_akas
region
movie_language
title_akas_types
title_akas_attributes
isOriginalTitle
reject_reason
DI_JobID
DI_Create_DT

dim_imdb_name_basics
name_sk
nconst
primaryName
birthYear
deathYear
name_imdb_url
DI_JobID
DI_Create_DT

dim_imdb_title_basics
title_sk
tconst
titleType_sk (FK)
movieId
imdbId
tmdbId
primaryTitle
originalTitle
isAdult
startYear
endYear
runtimeMinutes
title_imdb_url
DI_JobID
DI_Create_DT

dim_imdb_name_basics_primaryProfession
name_basics_primaryProfession_sk
name_sk (FK)
nconst
primaryProfession_sk (FK)
DI_JobID
DI_Create_DT

dim_imdb_name_basics_knownForTitles_rejects
knownForTitles_sk
title_sk (FK)
tconst
name_sk (FK)
nconst
reject_reason
DI_JobID
DI_Create_DT

dim_imdb_title_basics_genres
titles_genres_sk
title_sk (FK)
tconst
genres_sk (FK)
DI_JobID
DI_Create_DT

dim_imdb_genres
genres_sk
genres
DI_JobID
DI_Create_DT

dim_imdb_title_principals
title_principals_sk
title_sk (FK)
name_sk (FK)
title_ordering
job_category_sk (FK)
job_title
character_name
DI_JobID
DI_Create_DT

dim_imdb_title_principals_rejects
title_principals_sk
title_sk (FK)
tconst
name_sk (FK)
nconst
title_ordering
job_category_sk (FK)
job_category
job_title
character_name
reject_reason
DI_JobID
DI_Create_DT

fct_imdb_title_ratings
title_ratings_sk
title_sk (FK)
averageRating
numVotes
DI_JobID
DI_Create_DT

fct_imdb_brands_list
Brands_Lists_SK
brand_sk (FK)
release_Rank
title_sk (FK)
tconst
Release_Name
Lifetime_Gross
Max_Theaters
Opening_Gross
Open_Theaters
Release_Date
Distributor
Source_Data
DI_JobID
DI_Create_DT

fct_numbers_daily_box_office
daily_box_office_sk
title_sk (FK)
tconst
Movie_Name
Show_Date
Daily_Rank
Daily_Gross
Daily_Change_Pct
Weekly_Change_Pct
Number_of_Theaters
Gross_Per_Theater
Total_Gross
Number_of_Days
DI_JobID
DI_Create_DT

stg_imdb_titleType
titleType_sk
titleType
titleCategory
DI_JobID
DI_Create_DT

dim_imdb_title_episode
title_episode_sk
title_sk (FK)
parent_title_sk
tconst
parent_tconst
seasonNumber
episodeNumber
DI_JobID
DI_Create_DT

dim_imdb_title_akas
title_akas_SK
title_sk (FK)
tconst
title_ordering
title_akas
region
movie_language
title_akas_types
title_akas_attributes
isOriginalTitle
DI_JobID
DI_Create_DT

dim_imdb_job_category
job_category_sk
job_category
DI_JobID
DI_Create_DT

fct_movies_box_office_worldwide
movies_box_office_worldwide_sk
title_sk (FK)
tconst
BoxOffice_Rank
Worldwide_LifetimeGross
Domestic_LifetimeGross
Domestic_Pct
Foreign_LifetimeGross
Foreign_Pct
Release_Year
DI_JobID
DI_Create_DT

dim_imdb_title_crew_writers_rejects
writers_sk
title_sk (FK)
tconst
name_sk (FK)
nconst
reject_reason
DI_JobID
DI_Create_DT

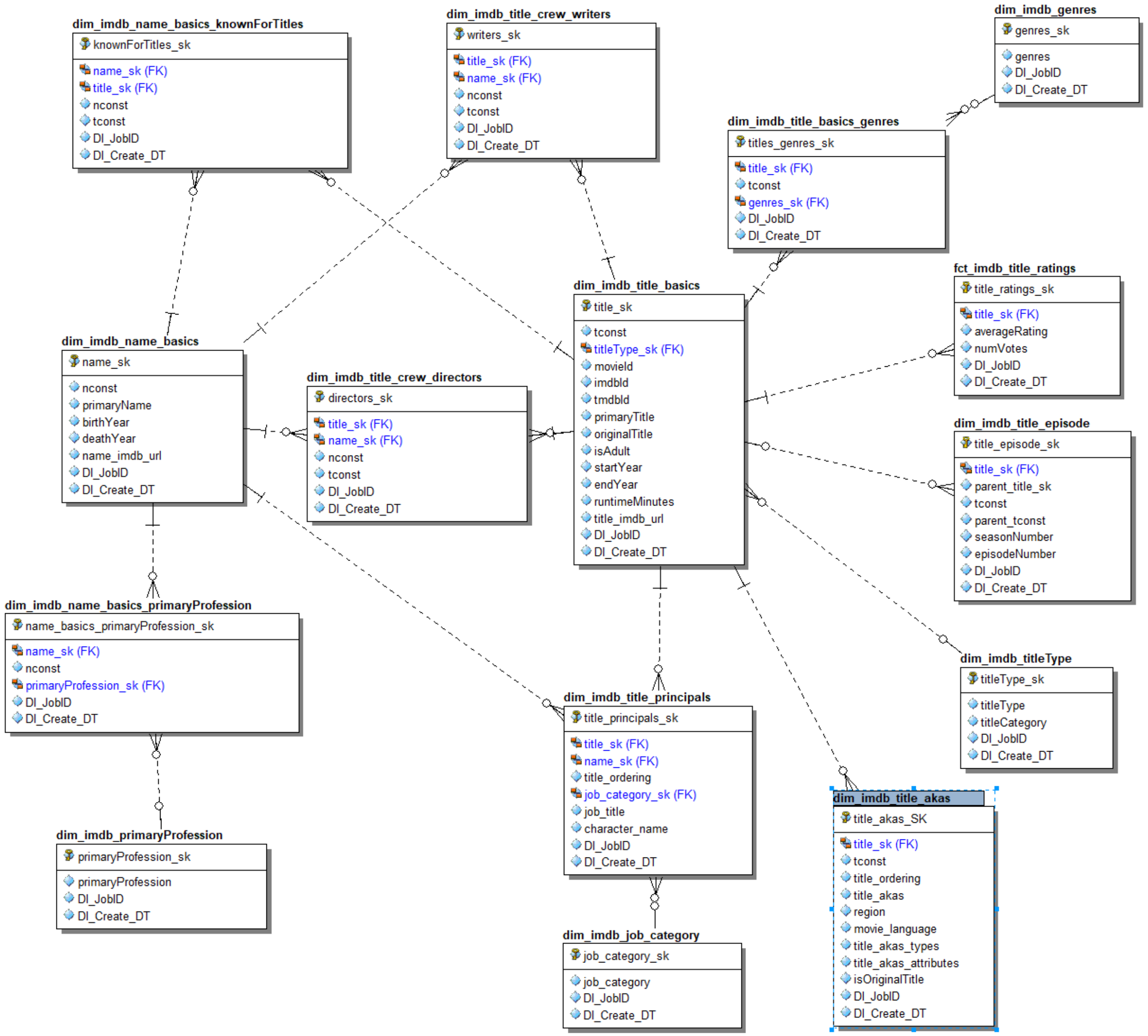
dim_imdb_title_episode_rejects
title_episode_sk
title_sk (FK)
parent_title_sk
tconst
parentTconst
seasonNumber
episodeNumber
reject_reason
DI_JobID
DI_Create_DT

dim_imdb_brand
Brand_sk
Brand_Name
DI_JobID
DI_Create_DT

fct_imdb_brands_gross
Brands_Gross_SK
brand_sk (FK)
Total_Revenue
Number_of_Releases
Top_Release
title_sk (FK)
Lifetime_Gross
DI_JobID
DI_Create_DT

ADA subproject (Core): Submodel – INT (imdb base)

UPDATED



dim_imdb_name_basics_knownForTitles
- knownForTitles_sk
- name_sk (FK)
- title_sk (FK)
- nconst
- tconst
- DI_JobID
- DI_Create_DT

dim_imdb_title_crew_writers
- writers_sk
- title_sk (FK)
- name_sk (FK)
- nconst
- tconst
- DI_JobID
- DI_Create_DT

dim_imdb_genres
- genres_sk
- genres
- DI_JobID
- DI_Create_DT

dim_imdb_title_basics_genres
- titles_genres_sk
- title_sk (FK)
- tconst
- genres_sk (FK)
- DI_JobID
- DI_Create_DT

dim_imdb_title_basics
- title_sk
- tconst
- titleType_sk (FK)
- movieId
- imdbId
- tmdbId
- primaryTitle
- originalTitle
- isAdult
- startYear
- endYear
- runtimeMinutes
- title_imdb_url
- DI_JobID
- DI_Create_DT

fct_imdb_title_ratings
- title_ratings_sk
- title_sk (FK)
- averageRating
- numVotes
- DI_JobID
- DI_Create_DT

dim_imdb_name_basics
- name_sk
- nconst
- primaryName
- birthYear
- deathYear
- name_imdb_url
- DI_JobID
- DI_Create_DT

dim_imdb_title_crew_directors
- directors_sk
- title_sk (FK)
- name_sk (FK)
- nconst
- tconst
- DI_JobID
- DI_Create_DT

dim_imdb_title_episode
- title_episode_sk
- title_sk (FK)
- parent_title_sk
- tconst
- parent_tconst
- seasonNumber
- episodeNumber
- DI_JobID
- DI_Create_DT

dim_imdb_name_basics_primaryProfession
- name_basics_primaryProfession_sk
- name_sk (FK)
- nconst
- primaryProfession_sk (FK)
- DI_JobID
- DI_Create_DT

dim_imdb_titleType
- titleType_sk
- titleType
- titleCategory
- DI_JobID
- DI_Create_DT

dim_imdb_title_principals
- title_principals_sk
- title_sk (FK)
- name_sk (FK)
- title_ordering
- job_category_sk (FK)
- job_title
- character_name
- DI_JobID
- DI_Create_DT

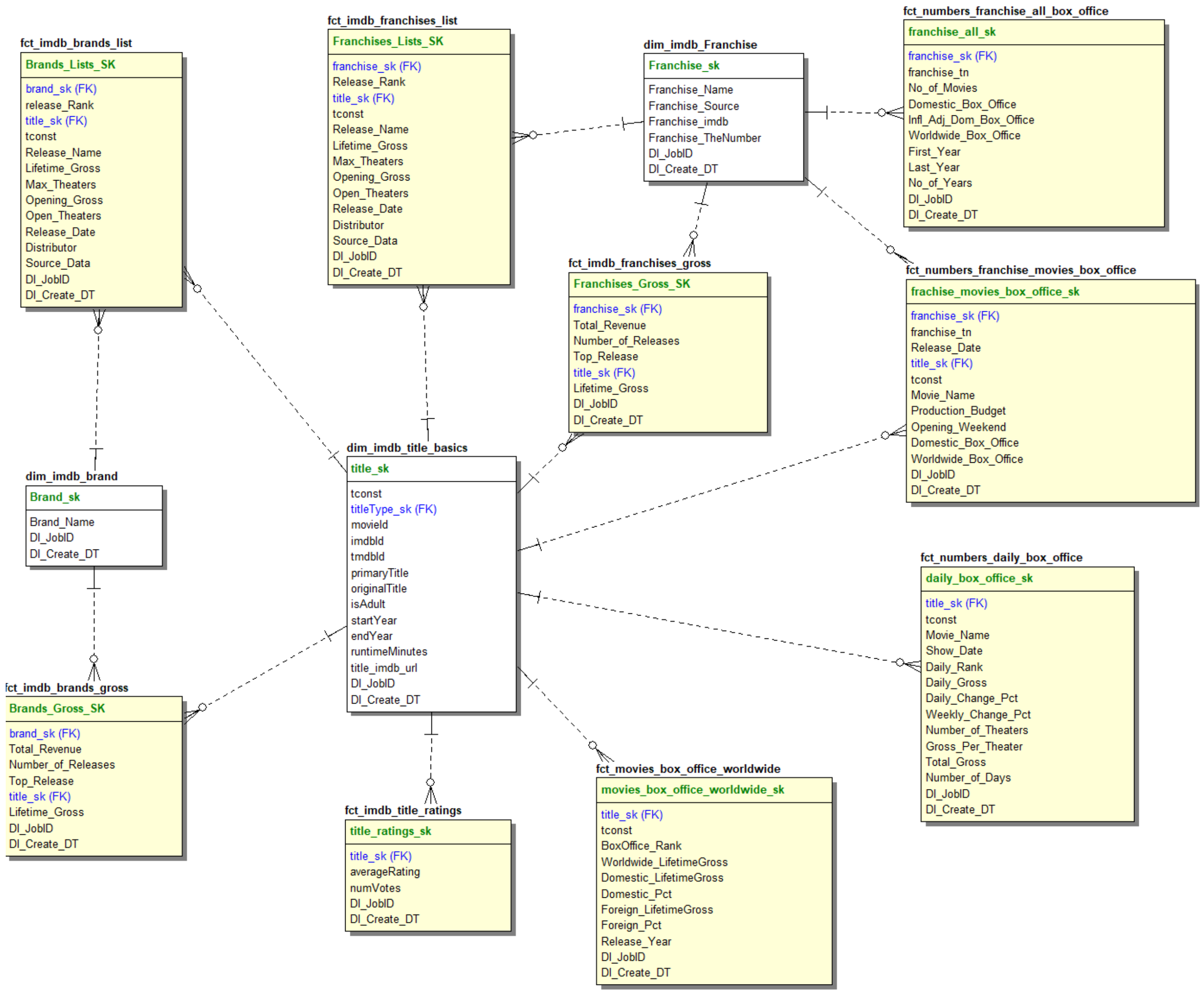
dim_imdb_title_akas
- title_akas_SK
- title_sk (FK)
- tconst
- title_ordering
- title_akas
- region
- movie_language
- title_akas_types
- title_akas_attributes
- isOriginalTitle
- DI_JobID
- DI_Create_DT

dim_imdb_primaryProfession
- primaryProfession_sk
- primaryProfession
- DI_JobID
- DI_Create_DT

dim_imdb_job_category
- job_category_sk
- job_category
- DI_JobID
- DI_Create_DT

ADA subproject (Core): Submodel – INT Facts

UPDATED



fct_imdb_brands_list
Brands_Lists_SK
brand_sk (FK)
release_Rank
title_sk (FK)
tconst
Release_Name
Lifetime_Gross
Max_Theaters
Opening_Gross
Open_Theaters
Release_Date
Distributor
Source_Data
DI_JobID
DI_Create_DT

fct_imdb_franchises_list
Franchises_Lists_SK
franchise_sk (FK)
Release_Rank
title_sk (FK)
tconst
Release_Name
Lifetime_Gross
Max_Theaters
Opening_Gross
Open_Theaters
Release_Date
Distributor
Source_Data
DI_JobID
DI_Create_DT

dim_imdb_Franchise
Franchise_sk
Franchise_Name
Franchise_Source
Franchise_imdb
Franchise_TheNumber
DI_JobID
DI_Create_DT

fct_numbers_franchise_all_box_office
franchise_all_sk
franchise_sk (FK)
franchise_tn
No_of_Movies
Domestic_Box_Office
Infl_Adj_Dom_Box_Office
Worldwide_Box_Office
First_Year
Last_Year
No_of_Years
DI_JobID
DI_Create_DT

fct_imdb_franchises_gross
Franchises_Gross_SK
franchise_sk (FK)
Total_Revenue
Number_of_Releases
Top_Release
title_sk (FK)
Lifetime_Gross
DI_JobID
DI_Create_DT

fct_numbers_franchise_movies_box_office
frachise_movies_box_office_sk
franchise_sk (FK)
franchise_tn
Release_Date
title_sk (FK)
tconst
Movie_Name
Production_Budget
Opening_Weekend
Domestic_Box_Office
Worldwide_Box_Office
DI_JobID
DI_Create_DT

dim_imdb_brand
Brand_sk
Brand_Name
DI_JobID
DI_Create_DT

dim_imdb_title_basics
title_sk
tconst
titleType_sk (FK)
movieId
imdbId
tmdbId
primaryTitle
originalTitle
isAdult
startYear
endYear
runtimeMinutes
title_imdb_url
DI_JobID
DI_Create_DT

fct_numbers_daily_box_office
daily_box_office_sk
title_sk (FK)
tconst
Movie_Name
Show_Date
Daily_Rank
Daily_Gross
Daily_Change_Pct
Weekly_Change_Pct
Number_of_Theaters
Gross_Per_Theater
Total_Gross
Number_of_Days
DI_JobID
DI_Create_DT

fct_imdb_brands_gross
Brands_Gross_SK
brand_sk (FK)
Total_Revenue
Number_of_Releases
Top_Release
title_sk (FK)
Lifetime_Gross
DI_JobID
DI_Create_DT

fct_imdb_title_ratings
title_ratings_sk
title_sk (FK)
averageRating
numVotes
DI_JobID
DI_Create_DT

fct_movies_box_office_worldwide
movies_box_office_worldwide_sk
title_sk (FK)
tconst
BoxOffice_Rank
Worldwide_LifetimeGross
Domestic_LifetimeGross
Domestic_Pct
Foreign_LifetimeGross
Foreign_Pct
Release_Year
DI_JobID
DI_Create_DT

IMDb Datasets: Box Office Revenues

UPDATED

- 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

Box Office Mojo: Box Office Revenues

UPDATED

- World-Wide Box Office All Time Top 1000 Movies
  - imdb_project - IMDb Mojo Box Office
  - 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

FileName	FileRows
World Wide Box Office All Time Top 1000.tsv	1000

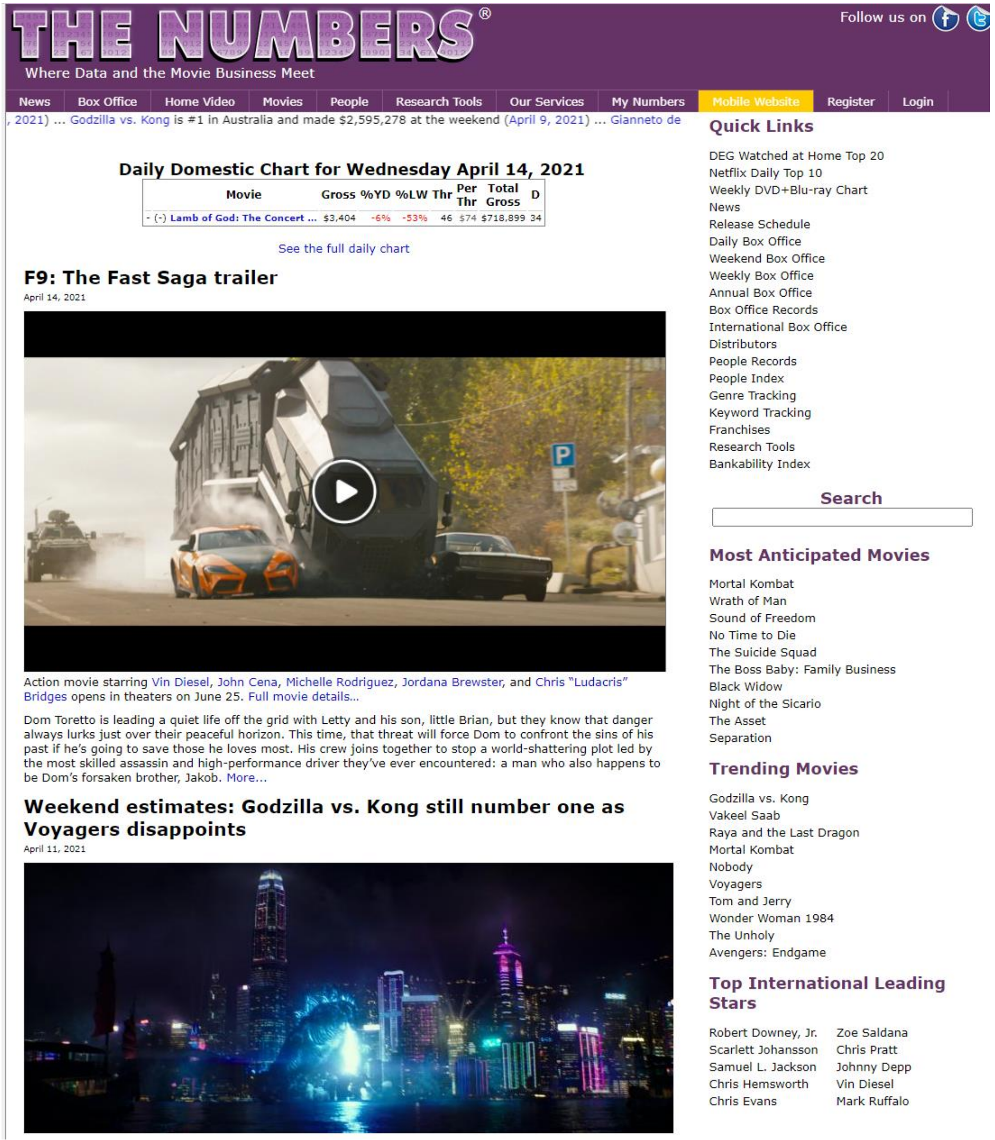
IMDb Datasets: Franchises & Brands

UPDATED

Category	FileName	FileRows
Brand	IMDb BoxOfficeMojo - Bad Robot.tsv	15
Brand	IMDb BoxOfficeMojo - bluesky.tsv	13
Brand	IMDb BoxOfficeMojo - blumhouse.tsv	47
Brand	IMDb BoxOfficeMojo - darkhorse.tsv	16
Brand	IMDb BoxOfficeMojo - dc.tsv	44
Brand	IMDb BoxOfficeMojo - dreamworks.tsv	37
Brand	IMDb BoxOfficeMojo - hasbro.tsv	16
Brand	IMDb BoxOfficeMojo - illumination.tsv	10
Brand	IMDb BoxOfficeMojo - Legendary.tsv	55
Brand	IMDb BoxOfficeMojo - lucas.tsv	38
Brand	IMDb BoxOfficeMojo - Marvel Comics.tsv	70
Brand	IMDb BoxOfficeMojo - MTV.tsv	36
Brand	IMDb BoxOfficeMojo - pixar.tsv	25
Brand	IMDb BoxOfficeMojo - platinum.tsv	17
Brand	IMDb BoxOfficeMojo - saturday alumni.tsv	30
Brand	IMDb BoxOfficeMojo - sony.tsv	21
Brand	IMDb BoxOfficeMojo - stephen.tsv	49
Brand	IMDb BoxOfficeMojo - Tim.tsv	8
Brand	IMDb BoxOfficeMojo - vertigo.tsv	40
Brand	IMDb BoxOfficeMojo - walden.tsv	38
Brand	IMDb BoxOfficeMojo - waltdisney.tsv	12

Category	FileName	FileRows
Franchise	IMDb BoxOfficeMojo - Franchise_ Avengers.tsv	4
Franchise	IMDb BoxOfficeMojo - Franchise_ Batman.tsv	19
Franchise	IMDb BoxOfficeMojo - Franchise_ Harry Potter.tsv	24
Franchise	IMDb BoxOfficeMojo - Franchise_ J.K. Rowling's Wizarding World.tsv	26
Franchise	IMDb BoxOfficeMojo - Franchise_ Marvel Cinematic Universe.tsv	37
Franchise	IMDb BoxOfficeMojo - Franchise_ Spider-Man.tsv	10
Franchise	IMDb BoxOfficeMojo - Franchise_ Star Wars.tsv	22
Franchise	IMDb BoxOfficeMojo - Franchise_ X-Men.tsv	14
Franchise	IMDb BoxOfficeMojo - Franchise_DC_Extended_Universe.tsv	13
Franchise	IMDb BoxOfficeMojo - Franchise_Disney_Live_Action_Reimaginings.tsv	17
Franchise	IMDb BoxOfficeMojo - Franchise_HungerGames.tsv	4
Franchise	IMDb BoxOfficeMojo - Franchise_James_Bond.tsv	26
Franchise	IMDb BoxOfficeMojo - Franchise_Jurassic_Park.tsv	8
Franchise	IMDb BoxOfficeMojo - Franchise_Middle_Earth.tsv	11
Franchise	IMDb BoxOfficeMojo - Franchise_Pirates.tsv	5
Franchise	IMDb BoxOfficeMojo - Franchise_Shrek.tsv	5
Franchise	IMDb BoxOfficeMojo - Franchise_Superman.tsv	8
Franchise	IMDb BoxOfficeMojo - Franchise_The_FastandtheFurious.tsv	11
Franchise	IMDb BoxOfficeMojo - Franchise_ToyStory.tsv	5
Franchise	IMDb BoxOfficeMojo - Franchise_Transformers.tsv	7

The Numbers



, 2021) ... Godzilla vs. Kong is #1 in Australia and made $2,595,278 at the weekend (April 9, 2021) ... Gianneto de

Daily Domestic Chart for Wednesday April 14, 2021

Movie	Gross	%YD	%LW	Thr	Per Thr	Total Gross	D
- (-) Lamb of God: The Concert ...	$3,404	-6%	-53%	46	$74	$718,899	34

See the full daily chart

F9: The Fast Saga trailer

April 14, 2021

Action movie starring Vin Diesel, John Cena, Michelle Rodriguez, Jordana Brewster, and Chris "Ludacris" Bridges opens in theaters on June 25. Full movie details...

Dom Toretto is leading a quiet life off the grid with Letty and his son, little Brian, but they know that danger always lurks just over their peaceful horizon. This time, that threat will force Dom to confront the sins of his past if he's going to save those he loves most. His crew joins together to stop a world-shattering plot led by the most skilled assassin and high-performance driver they've ever encountered: a man who also happens to be Dom's forsaken brother, Jakob. More...

Weekend estimates: Godzilla vs. Kong still number one as Voyagers disappoints

April 11, 2021

Quick Links

DEG Watched at Home Top 20
Netflix Daily Top 10
Weekly DVD+Blu-ray Chart
News
Release Schedule
Daily Box Office
Weekend Box Office
Weekly Box Office
Annual Box Office
Box Office Records
International Box Office
Distributors
People Records
People Index
Genre Tracking
Keyword Tracking
Franchises
Research Tools
Bankability Index

Search

Most Anticipated Movies

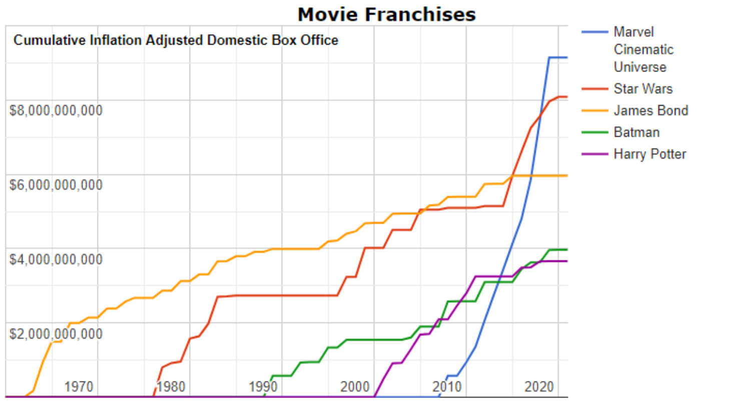
Mortal Kombat
Wrath of Man
Sound of Freedom
No Time to Die
The Suicide Squad
The Boss Baby: Family Business
Black Widow
Night of the Sicario
The Asset
Separation

Trending Movies

Godzilla vs. Kong
Vakeel Saab
Raya and the Last Dragon
Mortal Kombat
Nobody
Voyagers
Tom and Jerry
Wonder Woman 1984
The Unholy
Avengers: Endgame

Top International Leading Stars

Robert Downey, Jr.	Zoe Saldana
Scarlett Johansson	Chris Pratt
Samuel L. Jackson	Johnny Depp
Chris Hemsworth	Vin Diesel
Chris Evans	Mark Ruffalo



Search:

Franchise	No. of Movies	Domestic Box Office	Infl. Adj. Dom. Box Office	Worldwide Box Office	First Year	Last Year	No. of Years
Marvel Cinematic Universe	31	$8,545,968,433	$9,144,251,845	$22,557,315,232	2008	2022	14
Star Wars	15	$5,080,586,579	$8,082,302,972	$10,320,189,178	1977	2027	50
James Bond	27	$2,136,666,623	$5,963,688,920	$7,119,554,288	1963	2021	58
Batman	24	$2,783,757,158	$3,967,034,621	$6,040,885,610	1989	2022	33
Harry Potter	12	$2,786,938,291	$3,658,659,822	$9,215,022,636	2001	2022	21
Spider-Man	11	$2,711,115,724	$3,308,258,394	$7,223,015,335	2002	2020	18
X-Men	14	$2,458,465,265	$2,918,256,552	$6,075,357,352	2000	2020	20
Avengers	4	$2,619,552,260	$2,756,999,306	$7,752,758,277	2012	2019	7
Jurassic Park	6	$1,882,802,527	$2,720,632,476	$5,008,426,006	1993	2022	29
Star Trek	13	$1,400,952,879	$2,608,988,389	$2,266,907,732	1979	2016	37
Peter Jackson's Lord of the Rings	7	$1,852,304,517	$2,534,166,734	$5,847,175,428	2001	2014	13
DC Extended Universe	26	$2,193,942,393	$2,284,088,822	$5,636,366,972	2013	2023	10
Indiana Jones	5	$919,839,275	$2,074,807,903	$1,961,339,569	1981	2022	41
Superman	10	$1,152,698,852	$2,062,013,938	$2,553,272,239	1978	2020	42
Fast and the Furious	11	$1,690,752,038	$2,003,537,543	$5,893,692,187	2001	2022	21
Shrek	8	$1,419,598,493	$1,963,221,534	$3,545,629,858	2001	2022	21
Rocky	9	$792,039,875	$1,952,455,721	$1,513,943,877	1976	2022	46
Pirates of the Caribbean	5	$1,451,777,734	$1,922,575,299	$4,522,062,632	2003	2017	14
Transformers	7	$1,582,401,601	$1,873,340,793	$4,846,579,018	1986	2018	32
Toy Story	7	$1,318,120,746	$1,795,869,314	$3,054,457,147	1995	2022	27
Jaws	5	$420,895,891	$1,724,342,053	$737,573,891	1975	2020	45
Hunger Games	5	$1,451,538,526	$1,631,329,408	$2,958,353,344	2012	2015	3
Twilight	6	$1,365,922,346	$1,619,703,800	$3,317,470,739	2008	2012	4
Mission: Impossible	8	$1,154,492,507	$1,602,721,178	$3,577,423,206	1996	2023	27
Dark Knight Trilogy	3	$1,187,203,820	$1,489,433,835	$2,440,278,890	2005	2012	7
Peter Pan	6	$355,579,537	$1,475,963,881	$708,714,979	1953	2009	56
Planet of the Apes	9	$793,882,111	$1,382,261,890	$2,115,122,276	1968	2017	49
The Lion King	4	$965,423,326	$1,375,730,437	$2,640,596,802	1994	2019	25

The Numbers

UPDATED

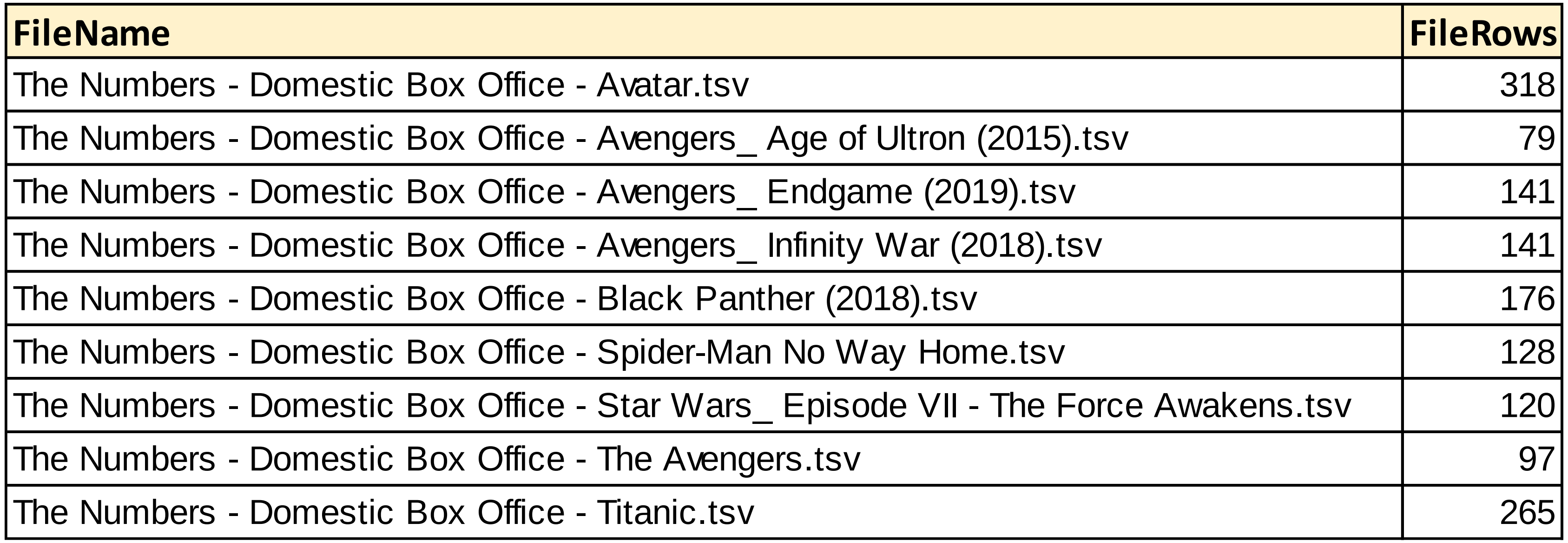
(Note: Not too excited about using this site but..)

- 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 (examples, see next page for list)
  - 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

The Numbers – Data Sources

UPDATED



FileName	FileRows
The Numbers - Domestic Box Office - Avatar.tsv	318
The Numbers - Domestic Box Office - Avengers_ Age of Ultron (2015).tsv	79
The Numbers - Domestic Box Office - Avengers_ Endgame (2019).tsv	141
The Numbers - Domestic Box Office - Avengers_ Infinity War (2018).tsv	141
The Numbers - Domestic Box Office - Black Panther (2018).tsv	176
The Numbers - Domestic Box Office - Spider-Man No Way Home.tsv	128
The Numbers - Domestic Box Office - Star Wars_ Episode VII - The Force Awakens.tsv	120
The Numbers - Domestic Box Office - The Avengers.tsv	97
The Numbers - Domestic Box Office - Titanic.tsv	265

IMDb project

Data Integration Rules

Data Integration – Standards

- Data Integration standards
  - All rows need to have DI_JobID and DI_CreateDT be filled in
  - All jobs must use Job Statistics Processing Joblets
    - DI Joblets and DI_CNTL database
    - Job runtimes will be documented using DI_CNTL database
  - All connections between Talend components need to be labeled i.e., no row1, row2, etc.
  - Orchestrator (or Master jobs) need to load all Staging, Integration and BI schemas
    - One job to load each of the above (Staging, Integration & each BI schema)
    - One job to run them all

- Best Practices reminders
  - Only use the columns needed when ingesting data
  - Trim data
  - Tweak batch/commit sizes, memory used, etc.

IMDb Project – Data Integration

- Load staging tables
  - Null values in data source files or tables need to result in SQL Server column Nulls
  - Check for file structural integrity (and reject anomalies)
  - Perform necessary data type conversions
- Perform data consistency & cleansing processes as appropriate
  - Do not allow duplicate data into ingestion schema from multiple data sources
  - Handle any file structural issues loading into staging tables
  - Incorporate rejection processing for integration schema
    - Rejection codes and descriptions need to be used
    - Except for file structural issues and eliminating duplicate data do not reject data loaded into stage tables

Data Integration Rules

- Do NOT use tUnite
- Do NOT use Bulk Load
- Do NOT physically combine any files for loading or cut & paste files together

Core IMDb project

Deliverables

IMDb Project Deliverables

- Ingest initial datasets into ingestion schema (staging tables)
- Integrate data into integration schema (dimensional data model)
- Load data into BI schema to better answer BI questions
- Perform data consistency & cleansing processes as appropriate
- Design and create BI visualizations answering business questions

Deliverables (Uploaded Files)

What to upload:

- Data Integration
  - List table row counts
  - Time to load each schema
- Talend
  - Export all your jobs with dependencies
  - Screenshots of your jobs’ workflows
  - Explain your data consistency, reject and structural integration processes
- BI
  - Screenshots of dashboards & explain purpose of each
  - PowerBI pbix files
  - Tableau twb files

Deliverables
(during project review)

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

IMDb Table Rows Counts

To Be Updated

TableName	T
stg_box_office_worldwide	To be filled in
stg_imdb_name_basics	
stg_imdb_name_basics_know	
stg_imdb_name_basics_p	
stg_imdb_title_akas	
stg_imdb_title_basic	
stg_imdb_title_ba	
stg_imdb_title_	
stg_imdb_title	
stg_imdb_tit	
stg_imdb_ti	
stg_imdb_	
stg_imdb_	
stg_iso_c	
stg_iso_la	
stg_ml_ge	
stg_ml_ge	
stg_ml_link	
stg_ml_mov	
stg_ml_rating	
stg_ml_tags	

_name_basics	To be led
dim_imdb_name_basics_knownForTitles	
m_imdb_name_basics_knownForTitles_rejects	
db_name_basics_primaryProfession	
primaryProfession	
akas	
rejects	
s	
di	
dim_im	
dim_imdb_	
dim_imdb_title_	
dim_imdb_title_episo	
dim_imdb_title_episode_	
dim_imdb_title_principals	
dim_imdb_title_principals_rejects	
dim_imdb_titleType	
dim_iso_country	
dim_iso_language	
imdb_title_ratings	
fct_ml_tags_	
fct_movies_box_office_worldwide	

TableName	Table_Rows
top1k_imdb_genres	To be filled in
imdb_job_category	
b_movie_roles	
ame_basics	
e_basics_knownForTitles	
basics_primaryProfession	
rofession	
bi genres	
bi_to ectors	
bi_top rs	
bi_top1k	
bi_top1k_	
bi_top1k_	
bi_top1k_	
bi_top1k_	
bi_top1k rldwide	